Preprint NCU/CCS-1999-1215; NCTS-99xxxx

Frequencies of Oligonucleotides as A Means of Characterizing Organisms

LIOAFU LUO[†‡],[1] FENG-MIN JI[†‡2] L.C. HSIEH[†] and H.C. LEE[†‡3]

[†]*Center for Complex Systems, National Central University, Chungli 320, Taiwan*
and
[‡]*National Center for Theoretical Sciences, P.O. Box 2-131, Hsinchu 300, Taiwan*

(Received December 17, 1999)

Frequencies of oligonucleotides are used to characterize nucleic sequences. The difference in frequency distributions of oligonucleotides n bases long in sequences is used to define an n-distance between sequences that are not aligned. Phylogenetic trees for 35 organisms, 19 bacteria, 9 archaeons and 7 eukaryotes, are constructed using n-distances, n=2 to 9, computed from the 16S/18S rRNA sequences of the organisms. The quality of the trees varies with the method of tree construction and generally improves with increasing n. The best trees are obtained at n=7 or 8. Not apparently useful phylogenetically are the 2- and 3-distances. On the best trees a number of features are correct, including the branching of the 35 organisms into the three kingdoms of Archaea, Eukarya and Bacteria, the deep branching of the thermotogales on the bacterial branch, the pairing of the closest sister organisms and the grouping of much of the close relatives. The thermotogales are conspicuously half bacterial and half archaeal.

[1]On leave from Department of Physics, Inner Mongolia University, Hohhot, 010021, China
[2]Present adress: The Institute of Theoretical Physics, Academia Sinica, Beijing 100080, China
[3]Send correspondence to: hclee@sansan.phy.ncu.edu.tw

INTRODUCTION

For some time now the standard measure of the molecular organismal evolutionary distance is based on differences in aligned molecular sequences [1, 2, 3]. Notwithstanding the impressive success of the approach, phylogenetic trees thus constructed based on different gene sequences may not be compatible [4]. Indeed, "no consistent organismal phylogeny has emerged from the many individual protein phylogenies so far produced [5]." One reason is that sequence mutation rates are not uniform across organisms. Another reason is the frequent occurrence of horizontal gene transfers that tend to confuse or mask the vertical lineages implied by trees [6, 7, 8]. At the very least, a consistent phylogeny will require the comparison of a selected set of several genes [9]. Technically, measuring distance by alignment makes heavy demands on computation when the compared sequences are long. In adition, an alignment based distance between to sequences is highly dependent on details of the alignment algorithm used and the sequence set chosen. In fact, a non-subjective multiple alignment of a set of distantly related long sequences is not feasible. There is therefore a need for supplementary definitions of evolutionary distances that do not require sequence alignment and are practical for comparing long sequences. It has been pointed out that short range correlations of nucleotides in a DNA sequence carry evolutionary information and they have been used to reconstruct partial evolutionary trees [10, 11, 12, 13]. We are therefore motivated to examine in some depth the utility of using oligonucleotide inventories to define organismal distance. By oligonucleotide inventory we mean all the oligonucleotides (up to a specific length) and their frequencies of occurrence in a DNA sequence. In this work we use only partial inventories by defining an n-distance that measures the difference in the frequency distributions of oligonucleotides of n bases long in two sequences (see the section "Methods" below). 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. It particular, it is independent on the size of the set of sequences being compared. For long sequences the computation time scales with sequence length. The method is similar in spirit to, but significantly different in practice from the so-called oligonucleotide cataloging method that pioneered the molecular approach to phylogeny [14] and which led to the discovery of Archaea, the third kingdom of organisms besides Bacteria and Eukarya [15].

In this work, for the purpose of testing the proposed method we apply it on the 16S rRNA (for archaeons and bacteria) and 18S rRNA (for eukaryotes) sequences for computing the n-distances. These genes are long known to have excellent clock-like behavior and were used to construct the first microbial phylogenetic tree for all prokaryotes [1, 16]. Thirty-five diverse organisms are considered in this study: 9 archaeons, 19 bacteria and 7 eukaryotes.

The well-known neighbor-joining (NJ) method, the unweighed pair-group mean arithmetic (UPGMA) method and the not so well-known fuzzy clustering (FC) method (see the section "Methods" below) are employed for tree construction. For simplicity we call a tree constructed from n-distances an n-tree. Ideally the benchmark against which one can test the quality of an n-tree could be a tree extracted from the Tree of Life

by removing all except the 35 organisms considered in this paper. We call this the life tree [17]. 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, called the alignment tree, using standard distances based on alignment of the 35 16S/18S rRNA sequences (while keeping in mind of the size dependence of the alignment-determined distance itself). The difference between the alingment and life trees provides us with a yardstick for measuring the difference between an n-tree and the life tree.

METHODS

Sequence data. The 35 organisms - 9 archaeons, 19 bacteria and 7 eukaryotes - studied in this work and the source 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. Where representatives of subclasses are missing in the first selection we take species whose genomes are not completely sequenced. 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.

The life tree. The life tree is obtained from existing consensus, alignment-based trees by removing from them all organisms or species not included in Table 1. Its three-kingdom topology is from [1]. Its Archaea and Bacteria branches are reconstructed from the prokaryotic tree of [17]. The Eukarya branch is from [2] and [18].

The rRNA alignment tree. The software package OMIGA 1.13 [19] was used to make multiple alignment of the 35 sequences listed in Table 1 and to construct alignment trees. No parts of any of the sequences were masked for the purpose of alignment. Two alignment trees were actually constructed, the 35-sequence tree based on multiple alignment of the entire 35-sequence set and the 3-kingdom tree where the three kingdoms were assumed and the three set of sequences - bacterial, archaeal and eukaryotic - were aligned separately.

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 bases a and b occurring sequentially in the sequence by p_{ab}. In general, if $\sigma = abc\cdots$ 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 legnth N, with increasing n the set $\{\sigma\}$ is an increasingly fine-grained characterization of a sequence. Given two sequences with sets of joint probabilities $\{p_\sigma\}$ and $\{p'_\sigma\}$, respectively, define the n-distance

$$E_n = \sum_\sigma |p_\sigma - p'_\sigma|, \qquad n = 1, 2, \ldots \tag{1}$$

between the two sequences. An n-distance is well defined for sequences that are of different lengths and are not aligned. By repeated application of relations such as

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

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

$$E_n \leq E_{n+1}, \qquad n = 1, 2, \cdots \tag{3}$$

The limit of the increasing series $E_1, E_2, \cdots$, 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 $i, j, k, \ldots$, 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 computed from the two sequences representing the organisms i and j. 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 and the length for the aligned sequence is L, then

$$E_1 = \frac{1}{L}\sum_a \left| n_{a\to\not a} - n_{\not a\to a} \right| \leq \frac{1}{L}\sum_a \left(n_{a\to\not a} + n_{\not a\to a} \right) \equiv 2M < 2 \tag{4}$$

where M is the total number of single mutations divided by L, namely the fraction of positions in which the two sequences differ; $n_{a\to\not a}$ is the number of incidents where a base a in the first sequence is either changed to another base ($\not a$) or is missing in the second sequence; and $n_{\not a\to a}$ is the number of incidents where either a blank or a base that is not a in the first sequence is changed to a in the second sequence. (If the two sequences are of different lengths, then blanks generated in the alignment must be counted as a fifth kind of nucleotides for the first equality in Eq. (4) to hold.) The parameter M is the conventional definition of evolution 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, 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. For instance, if two sequences are identical except that one has an extra base, then the distance between the two sequences is at most n/N, where N is the length of the shorter sequence. Unlike M, which has a first order relation with the difference in evolution time (assuming the rate of mutation is constant, which may not be true), there is not an intuitively obvious relation between an n-distance and evolution time, other than that the two are expected to be positively correlated. Neither is it obvious whether M or some n-distance is the closer approximation to the true evolution time. Suppose an approximate scaling relation between an n-distance and evolution time can be established. Then, because no consideration is given to related but non-identical oligomers, an n-distance tends to overestimate the (scaled) true evolutionary distance [14].

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 Table 1, where the element d_{ij} is the n-distance between the rRNA sequences of the i^{th} and j^{th} organisms. The sequences are not aligned. Dendograms, or n-trees, are then constructed from the distance matrix using the UPGMA, FC and NJ methods.

The UPGMA and NJ methods. These two well-known methods (and others) are described in, for instance, the book by Li [3]. For tree construction and plotting we use the software package PHYLIP version 3.5c by Felsenstein [20]. A rooted tree is obtained by identifying an obvious out-group on an unrooted tree.

The fuzzy clustering method. Unlike the NJ and UPGMA methods, 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. A tree constructed by partition is built from the root up, namely the deepest or earliest branching is identified first, followed by the next deepest, and so on. This contrasts with the NJ and UPGMA methods in which the latest branching tips are identified first. Given a distance matrix d of N organisms the method is composed of three steps:

(1) Define a similarity matrix S (called fuzzy similarity matrix in fuzzy set theory) by $S = I - d/2$, where I is a matrix whose every element is 1. Because d is symmetric ($d_{ij} = d_{ji}$) with vanishing diagonal elements, S is symmetric and reflexive (S_{jj} =1). S_{ij} measures the closeness of the two objects i and j.

(2) Construct the fuzzy equivalence matrix T. The matrix T is defined as $T \equiv S \circ S \circ \cdots \circ S = S^{\circ(N-1)}$, where in the multiplication "$\circ$" of matrices the usual sum is replaced by the operation max (maximum value in a set) and usual product by the operation min (minimum value in a set). T is not only symmetric and reflexive but also transitive: $(T \circ T)_{ij} \leq T_{ij}$. Hence it is a matrix of equivalence relations.

(3) Construct a partition tree from T. Because T is a matrix of equivalence relations, it can be used to construct a unique rooted partition tree using the alpha-cut method [21].

RESULTS AND DISCUSSIONS

Oligonucleotide catalog, oligonucleotide inventory and n-distance. The method of using oligonucleotides to characterize a sequence is not new. Before the time when rRNAs could be completely sequenced they were characterized by their oligonucleotide catalogs, each produced by digestion of a sequence with a cleaving enzyme such as ribonuclease T_1. Phylogenetic studies based on oligonucleotide cataloging in fact led to the discovery of the Archaea kingdom [14]. With the availability of complete nucleic sequences complete oligonucleotide inventories can now be generated *in silico*. There are several important differences between using an oligonucleotide inventory to characterize a sequence and using an oligonucleotide catalog produced by cleaving *in vitro*. (a) Only a partial catalog of oligomers occurring in a sequence is actually generated through cleaving by nucleases. (b) A catalog produced through cleaving does not give information on the frequencies of occurrence of the oligomers in the catalog. (c) In the case of 16S/18S rRNA sequences that are 1500 to 1800 bases long, the lack of information on frequency meant that only oligomers at least 6 bases long should be included in the partial catalog [14], for only then would those oligomers appearing in the catalog have (on average) an actual frequency of one. (d) As an object that characterizes a sequence, a catalog depends on the cleaving nuclease as well as on the sequence itself. (e) On the other hand, whereas a catalog obtained by cleaving could include long (up to n=20 or so) oligomers just as easily (and naturally) as short ones, the computer storage requirement for constructing *in silico* the complete inventory of oligomers of length n grows as 4^n.

In this work, because of (e) above and because we are only testing the method we set $n \leq 9$. Also we do not actually use the entire inventory all at once, but rather classify it according to n and use each class separately. We leave the approach of using the complete inventories to determine distances to a future study.

The three kingdoms. Fig. 1 shows the life tree [1, 2, 17, 18] including only the 35 organisms listed in Table 1. The early divergence of the Tree of Life had been problematic [22, 23, 24], but seems to be settling on a consensus branching pattern (Bacteria(Archaea,Eukarya)) for the three kingdoms [9, 25]. All the alignment and best n-tree favor the branching (Eukarya(Bacteria,Archaea)) instead. Fig. 2 gives the 35-sequence alignment tree. Fig. 3 shows the best, or n=8, tree constructed using the UPGMA method. Fig. 4 shows the 2- and 8-trees constructed using the FC method; the latter is the best FC tree. Fig. 5 shows the n= 2 to 7 unrooted trees constructed using the NJ method. Fig. 6 shows the best, or n=7, NJ tree. Table 2 gives separately the branching patterns of the eukaryotic, archaeal and bacterial kingdoms on the life tree, the 35-sequence and 3-kingdom alignment trees and the best n-trees; not all levels of branching of Bacteria are indicated.

The life and alignment trees. The lineage of an organism as expressed on a phylogenetic tree depends on the size of the tree, because it is sensitive to what other organisms are included. When inter-organism distance is based on multiple sequence alignment, the sensitivity is magnified because the distance itself depends on the included organism set. This is the main reason why the crown eukaryotes branch as (fungi(plants(animals))) in [1], where the organism set is relatively small, but branch as (plants(fungi(animals))) in [18], where the set is much larger. For the Eukarya kingdom of the life tree (Fig. 1) we use the branching given in [18].

For the same reason mentioned above details of an alignment tree depend on whether the tree is constructed based on a 35-sequence alignment (the 35-sequence tree, Fig. 2) or whether sequences in the three kingdoms are aligned separately (the 3-kingdom tree). The two sets of branchings are both given in Table 2.

Both versions of the alignment tree correctly branch into three kingdoms and have all the very close relatives correctly paired off, but they differ from each other and from the life tree in detail. On both versions, in Eukarya the *C. elegans* branches too early while the plants (*G. max* and *S. tuberosum*) and yeast branch too late. In Archaea the crenarchaea (*A. pernix*, *T. tenax* and *S. solfataricus*) are correctly separated from the euryarchaeotes, but the topology of the latter differs from its counterpart on the life tree. In particular the positions of thermococcus *P. horikoshii* tend to branch too early and the extreme halophile *H. volcanii* tend to branch too late.

The main difference between the 35-sequence and 3-kingdom alignment trees occurs in the placement of the green non-sulfur *H. aurantiacus* on their respective bacterial branches. Both trees correctly have the thermophiles *A. aeolicus* and *T. maritima* branch off first. On the 3-kingdom tree this is followed by *H. aurantiacus* and the radioresistant micrococci *D. radiopugans* consistent with the life tree. On the 35-sequence tree *D. radiopugans* is mixed with the gram-positives (*B. subtilis* and *M. tuberculosis*) and cynobacteria (*Synechococcus sp.*) late on the branch instead. There are irregulari-

ties common to both versions. A conspicuous example is the failure of the mycoplasmas (*M. genitalium* and *M. pneumoniae*) to group with the gram-positives. Also the flexibacteria (*F. heparinum* and *Ch. limicola*) are placed close to the proteobacteria (*E. coli*, *H. influenzae*, *H. pylori* and *R. prowazekii*) instead of the spirocheoets (*B. burgdorferi* and *T. pallidum*). The unusualness of the mycoplasmas (or rather their rRNAs) was pointed out by Woese some time ago [1]: they show a tendency to vary otherwise conserved positions in rRNAs more readily than do other bacteria. The tendency may be attributed to the fact that their genomes are small, so they can more easily develop elevated mutation rates.

General features of the n-distance trees. An n-distance between two sequences is, unlike distances based on sequence alignment, independent of the size of the sequence set. The topology of n-distance based dendograms depends on the method of construction as well as on n. On the NJ and UPGMA (as well as the alignment) trees branching and branch length are both meaningful, but on the latter all the tips are equidistant from the root, which is equivalent to assuming a constant molecular clock. On the FC tree only branching has significance. The general trend is that the quality of the n-trees improves (*i.e.*, it becomes more similar to the alignment tree) with increasing n when $n \leq 6$, and reaches a plateau at n=7 or 8. Most of Archaea and most of Eukarya form their own separate groups on trees with $n \geq 4$ but in one instance (the FC tree) the three kingdoms are recognizable even on the 2-tree. Typically when Archaea is not completely separated from the rest it is missing *M. thermoautotrophicum* and *H. volcanii* and is contaminated with the thermotogales *A. aeolicus* and *T. maritima*. For $n \leq 6$ the UPGMA and FC trees are better in separating the three kingdoms but for $n \geq 7$ NJ gives the better trees. For all methods the best n-trees are obtained at n=7 or 8.

On the whole the best n-trees are closer to the 35-sequence alignment tree than to the 3-kingdom alignment tree for the Archaea and Eukayra branches. For the Bacteria branch the situation is less clear cut. The Bacteria branch on the overall best n-tree, the NJ 7-tree, is closer to its counterpart on the 3-kingdom tree. Not surprisingly, some of the inconsistencies of the alignment trees relative to the life tree are shared by the best n-trees. The most conspicuous ones are: having *C. elegans* as the deepest branching eukaryote; the exchange of position between *H. volcanii* and *P. horikoshii*; the separation of *D. radiopugans* from *H. aurantiacus*; the misplacement of the mycoplasmas.

A robust feature that persist on all the best (and a number of not-so-best) trees is the correct pairing of eight sets of the closest relatives (called sisters below): *H. sapiens* and *M. musculus*, *S. tuberosum* and *G. max*, *M. fervidus* and *M. thermoautotrophicum*, *M. genitalium* and *M. pneumoniae*, *E. coli* and *H. influenzae*, *Ch. trachomatis* and *Ch. pneumoniae*, *B. burgdorferi* and *T. pallidum*, *A. aeolicus* and *T. maritima*. The closely related crenarchaeotes *A. pernix* and *T. tenax* are correctly paired on some n-trees, but not on the best ones. In contrast, close relatives (on the life tree) *D. radiopugans* and *H. aurantiacus*, and the not-so close relatives *A. fulgidus* and *H. volcanii*, and *F. heparinum* and *Ch. limicola* are never paired on any of the n-trees. Only the last two are correctly paired on the alignment trees.

The UPGMA trees. The UPGMA trees gets closer monotonically to the alignment

tree with increasing n, up to n=8. Archaea is an identifiable out-group on all the trees. But for $n \leq 6$ it has *M. thermoautotrophicum* and *H. volcanii* missing and is contaminated with the thermotogales *A. aeolicus* and *T. maritima* from Bacteria. The seven eukaryotes are separated cleanly from Bacteria on trees with $n \geq 4$. That is, on these trees the three kingdoms are recognizable. The topology is (Archaea, (Bacteria, Eukarya)) for n= 4, 5 and 6 but switches to ((Archaea, Bacteria), Eukarya) at n=7, when the three kingdoms become cleanly separated. The topology of the tree changes only slightly thereafter. Fig. 3 shows the best UPGMA tree with n=8.

Aside for giving the three kingdoms correctly, the 8-tree has a number of correct details: in Archaea the three crenarchaeotes (*A. pernix*, *T. tenax* and *S. solfataricus*) are grouped apart from the five euryarchaeotes; in Eukarya the animals (except *C. elegans*) are set apart from the plants and yeast; the eight pairs of sisters are correctly paired.

The 8-tree is inferior to the alignment tree in a number of the features: in Archaea *S. solfataricus* instead of *T. tenax* is paired with *A. pernix*; in Bacteria the pair of mycoplasmas *M. genitalium* and *M. pneumoniae* branch most deeply, before *H. aurantiacus*; the thermotogales branch after *H. aurantiacus*; aside from the pairing of the sisters, the topology of the Bacteria branch is basically inconsistent with the life tree.

The FC trees. As a function of n the FC trees differ from the UPGMA trees. A unique feature is that the three kingdoms are more or less recognizable on the 2-tree (see Fig. 4). This is unexpected because the 2-distance is defined in terms of only 16 pieces of data on each sequence. This feature however disappears at n=3, re-emerges at n=4, disappears again at n=6 and re-emerges again to stay at n=7. Archaea is an identifiable group on all the trees. But *M. thermoautotrophicum* is missing when $n \leq 3$, *H. volcanii* is missing when $n \leq 4$ and *A. aeolicus* and *T. maritima* are included when $n \leq 6$. Eukarya forms a group by itself on all trees except when n=3. The out-group is Archaea when $n \leq 3$ but switches to being Eukarya when $n \geq 4$. The three kingdoms are cleanly formed for the first time at n=7. The thermotogales form the deepest branch in Archaea when n=5 and 6 but switches suddenly to branching quite late in Bacteria when $n \geq 7$. Riding over such fluctuating dependence on n is the general trend that the quality of the tree improves with increasing n. The best FC tree is n=8 with n=7 being a close second.

On the 8-tree (see Fig. 4) Eukarya is the same as that on the 35-sequence alignment tree. Archaea is similar but *A. pernix* is paired with *S. solfataricus* instead of *T. tenax*. In Bacteria there is much that differs from the alignment trees. Most glaring is the deep branching of the mycoplasmas and the relatively late branching of the thermotogales. Overall the topology of Bacteria on the n=8 FC tree is still poor, but is slightly better than that on the best (n=8) UPGMA tree.

The NJ trees. The NJ trees improves monotonically with increasing n, up to n=7, but the n-dependence of the trees differ qualitatively from those of the UPGMA and FC trees. At n=4 Eukarya clade and Archaea roughly form a group, but the three kingdoms are not yet formed. The unrooted trees for n= 2 to 7 are shown in Fig. 5. As in the UPGMA and FC trees, for smaller values of n the thermotogales *A. aeolicus*

and *T. maritima* tend to be grouped with Archaea (for $n \leq 5$) and the archaeons *M. thermoautotrophicum* and *H. volcanii* tend to be grouped with Bacteria (for $n \leq 4$). However unlike the UPGMA trees the eukaryotes are not disentangled from the bacteria until n=7, when the tree undergoes a phase transition and the three kingdoms take shape cleanly.

The best overall NJ tree, obtained when n=7 and rooted with Eukarya as the out-group, is shown in Fig. 6. On this tree Eukarya is identical to its counterpart on the 35-sequence alignment tree and the best (n=8) UPGMA and FC trees and Archaea is identical to its counterpart on the best UPGMA tree.

The Bacteria branch, although still not quite right, is much closer to Bacteria on the 3-kingdom alignment tree than its counterparts on the best UPGMA and FC trees. In particular the thermotogales branch deepest, followed by *H. aurantiacus*, and *D. radiopugans* closely behind. The four proteobacteria are also in a group (not so on the best UPGMA and FC trees). On the other hand, *F. heparinum* and *Ch. limicola* are still widely separated but should not be.

The eight pairs of sisters. If two sequences have a very high degree of homology, then their n-distances will be small for every n. This seems to be the case for the four pairs of organisms, the mammals *H. sapiens* and *M. musculus*, the plants *G. max* and *S. tuberosum*, the chlamydiae *Ch. trachomatis* and *Ch. pneumoniae*, and the mycoplasmas *M. genitalium* and *M. pneumoniae* whose aligned sequences are 98%, 95%, 93% and 97% identical, respectively. On every tree with $n > 2$, regardless of the method used to construct it, each of the four pairs are mutually the closest neighbors.

Four other closely related pairs, the euryarchaeotes *M. thermoautotrophicum* and *M. fervidus*, the protoebacteria *E. coli* and *H. influenzae*, the spirochetes *B. burgdorferi* and *T. pallidum*, and the thermotogales *A. aeolicus* and *T. maritima* are not always paired on the constructed trees but are so on all the best trees ($n \geq 7$). Their aligned sequences are 89%, 86%, 80% and 77% identical, respectively.

The archaeons. The three crenarchaeotes, *A. pernix*, *T. tenax* and *S. solfataricus*, stay close on most trees. On the n=2 to 4 trees *A. pernix* and *T. tenax* are correctly nearest paired neighbors but *S. solfataricus* is distant. On n=5 and 6 trees a pattern consistent with the life tree ((*T. tenax*, *A. pernix*) *S. solfataricus*) is obtained but on the best trees (n=7 to 9) a slightly different pattern (*T. tenax* (*A. pernix S. solfataricus*)) is obtained.

On the $n \leq 4$ trees *M. thermoautotrophicum* is incorrectly placed in Bacteria. It is placed in Archaea for $n \geq 5$. On the $n \geq 6$ trees it is correctly paired with *M. fervidus* as its nearest neighbor.

On the $n \leq 4$ trees *H. volcanii*, like *M. thermoautotrophicum*, is incorrectly placed in Bacteria and *P. horikoshii* is incorrectly grouped with the crenarchaeotes. On trees with $n \geq 6$ *H. volcanii* is either an out-group of Archaea or Euarchaeota while *P. horikoshii* is paired with *A. fulgidus*.

On the best NJ, FC and UPGMA trees (n= 7 or 8) the archaeons correctly divide into a group of three crenarchaeotes and a second group of six euryarchaeotes, as they do on the 35-sequence alignment tree. But in all cases the positions of *H. volcanii* and *P. horikoshii* are inverted (relative to the life tree). As far as the archaeons are concerned

the 7- and 8-distances, as measures of evolutionary distance, are just slightly inferior to distances determined by sequence alignment.

The eukaryotes. On the n-trees the eukaryotes form a group by themselves for good starting at n=4 and, depending on the method of tree construction, take their final branching pattern at n=7 or 8. On the best trees the branching pattern is identical to that on the 35-sequence alignment tree: (nematode((fly(mammals))(yeast(plants)))), as opposed to the pattern on the life tree (plants(yeast(animals))) [9, 18]. Thus for the set of test eukaryotes the 7- and 8-distances are just as good as distances determined by sequence alignment.

The bacteria. On the lower n-trees the bacteria either have some archaeons and/or eukaryotes mixed in or have Eukarya as a division. They form a group by themselves for good starting at n=7. As seen in Table 2, the branching pattern of the bacteria is very sensitive to n (see Fig. 5) and the tree construction method. The last column in the table gives the number of moves needed to bring a branching pattern into agreement to that on the life tree. According to this measure the NJ tree and the 35-sequence alignment tree are equidistance from the life tree, the 3-kingdom tree is closer and the FC and UPGMA trees are farther.

The thermotogales. Early rRNA-based phylogenies have designated the thermotogales *A. aeolicus* and *T. maritima*, both hyperthermophiles, to be among the deepest branching bacteria [26, 27], but this placement is not often supported by other protein-based phylogenies [4]. With the complete sequencing of the genomes the lineages of the two organisms are under re-examination. Although the bacterial origin of *A. aeolicus* and *T. maritima* is not challenged, preliminary comparative analyses of a number of proteins in *T. maritima* do not yield a statistically significant placement of its lineage [6]. As well comparative analyses of the proteins of *A. aeolicus* do not yield a clear picture, but offer little support for its rRNA-based position [7]. These analyses indicate that both hyperthermophiles share a common ancestor with Bacteria for a majority of genes involved with housekeeping functions such as transcription, translation, DNA replication and cell division, but inherited from the ancestor of Archaea about half of their genes involved with metabolic functions [6, 7]. The mixed heritage of the hyperthermophiles has been taken to be evidence of extensive horizontal gene transfers between Archaea and Bacteria.

On the n-trees, from n=3 on, and disregarding the two archaeons *M. thermoautotrophicum* and *H. volcanii* that are often stranded (see discussion above), the two thermotogales always form a distinct group separate from everything else. Interestingly the ambiguous nature of the lineage of the thermotogales is evident on our constructed trees even as they are based on a single rRNA. On all three types of constructed trees they are placed in Bacteria when $n \leq 6$, often being an out-group, and go over to Archaea when $n \geq 7$. This pattern is typically revealed on the unrooted NJ trees (Fig. 5). However, only on the best NJ tree is the pair the deepest branching bacterial (see Table 2).

The mycoplasmas. Owing to very high similarity the pair of mycoplasmas *M. geni-*

talium and *M. pneumoniae* are never separated on any of the trees. On the other hand their affinity to any other group is highly dependent on n and on the method of tree construction, as is partly evident from the way the pair jumps around on the bacterial branches shown in Table 2. This suggests that the n dependence of the distance between the pair follows a pattern that deviates from the average pattern, which is consistent with the fact that on the mycoplasmal rRNA many otherwise conserved positions are mutated [1].

CONCLUSION

Frequencies of oligonucleotides of n=2 to 9 bases long were used to define n-distances between the 16S/18S rRNA sequences of 35 organisms, phylogenetic trees for these organisms were constructed using several distance methods, and the trees were compared with benchmark trees based on sequence alignment. The constructed trees display a strong dependence on n and on the construction method. The property of the tree generally improves with increasing n. The overall pattern of the n-dependence is: recognizable Archaea starting from n=2, formation of Eukarya as a separate group from n=4, and formation of the three kingdoms from n=7. The 3-distance does not in anyway stand out. For each method used, the 7- and 8-distances are found to be best able to represent evolutionary distance. For the set of 35 organisms chosen for this work, the branching pattern of the bacterial group is the key for ranking the trees. The n=7 tree constructed with the neighbor-joining (NJ) method is the best n-tree. In comparison with the other n-trees it has a significantly higher degree of similarity with the alignment trees. When the life tree is used as the benchmark, the n=7 NJ tree is as good as the 35-sequence alignment tree and only slightly inferior to the 3-kingdom alignment tree. An interesting finding is the placing of the two thermotogales *A. aeolicus* and *T. maritima*, which display a persistent tendency to be "half bacterial half archaeal". In contrast, the mycoplasmas display an especially volatile dependence on tree construction methods, but on none among the n-trees and alignment trees is their placing consistent with the life tree. In conclusion, the 2- and 3-distances are not useful measures for phylogeny; the 4- 5- and 6-distances can be used to place an organism in its correct kingdom in most cases and in its correct division is some cases; the 7- 8- and 9-distances can be expected to place an organism in its correct kingdom always and to produce a reasonable phylogenetic tree. Since any given set of n-distances constitutes only one part of the complete inventory of oligomers (up to a certain length) in a sequence, it remains to be seen if a larger portion of the inventory, or even the complete inventory, can do a better job.

This work is partly supported by grants NSC 88-2816-M007-0005-6 to LLF and NSC 89-M-2112-008-019 to HCL from the National Science Council (ROC).

References

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

[2] Cavalier-Smith, T. Kingdom protozoa and its 18 phyla. *Microbiol. Rev.* **57**, 953-994 (1993).

[3] Li, W.H. *Molecular Evolution.* (Sinauer Associates, 1997).

[4] Daldauf, S.L., Palmer, J.D. & Doolittle, W.F. The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. *Proc. Natl. Acad. Sci. USA* **93**, 7749-7754 (1996).

[5] Woese, C.R. The universal ancestor. *Proc. Natl. Acad. Sci. USA* **95**, 6854-6859 (1997).

[6] Deckert, G. *et al.* The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. *Nature* **392**, 335-358 (1998).

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

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

[9] Feng, D.F., Cho, G. & Doolittle, R.F. Determining divergence times with a protein clock: Update and reevaluation. *Proc. Natl. Acad. Sci. USA* **94**, 13028-13033 (1997).

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

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

[12] Karlin, S., Mrazek, J. and Campbell, A.M. Compositional biases of bacterial genomes and evolutionary implications. *J. Bacteriology* **179**, 3899-3913 (1997).

[13] Luo, L.F., *et al.* Statistical correlation of nucleotides in a DNA sequence. *Phys. Rev. E* **58**, 861-871 (1998).

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

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

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

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

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

[19] Calvet, J.P. Comprehensive sequence analysis: OMIGA 1.1. *Science* **282** 1057-1058 (1998).

[20] Felsenstein, J. 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].

[21] Zadeh, L. Similarity relation and fuzzy orderings. *Info. Sci.* **3**, 177-206 (1971).

[22] Lake, J.A. Prokaryotes and archaebacteria are not monophyletic. *Cold Spring Harbor Symposium on Quantitative Biology*, **52**, 839-846 (1987)

[23] Moores, A.O. & Redfield, R.J. Digging up the roots of life. *Nature* **379**, 587-588 (1996).

[24] Doolittle, R.F. *et al.* Determining divergence times of the major kingdoms of living organisms with a protein clock. *Science* **271**, 470-477 (1996).

[25] Doolittle, F. Fun with genealogy. *Proc. Natl. Acad. Sci. USA* **94**, 12751-12753 (1997).

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

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

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: Comparison of first few levels of branchings of Bacteria on the various trees. Organisms are represented by codes given in Table 1. The "Tree of Life" is called the life tree in the text. The last column gives the number of moves needed to bring the branching pattern into agreement with the life tree.

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	

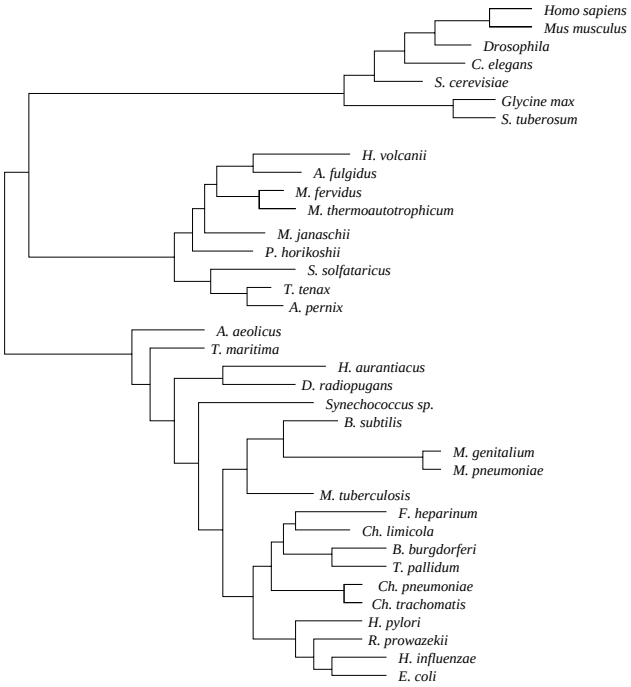


Figure 1: Tree of Life from 16S/18S rRNA alignment, with all except the 35 organisms listed in Table 1 removed. The Archaea and Bacteria kingdoms are reconstructed from [17], and the Eukarya kingdom is from [2, 18]. Branch lengths are only approximately to scale.

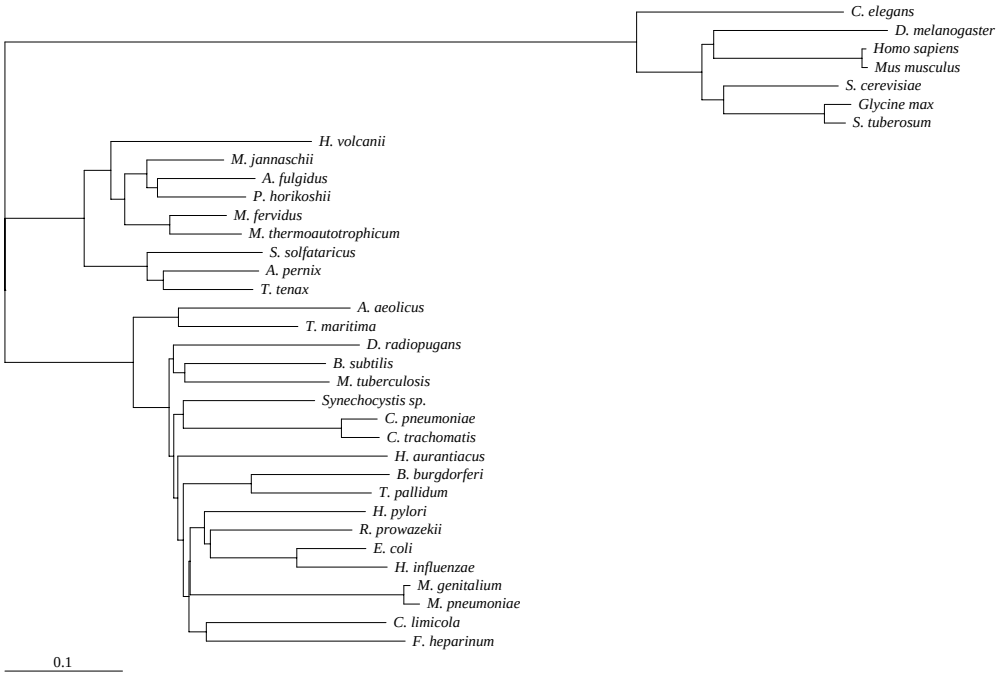


Figure 2: Tree from alignment of 16S/18S rRNA sequences of 35 organisms listed in Table 1 with Eukarya as out-group.

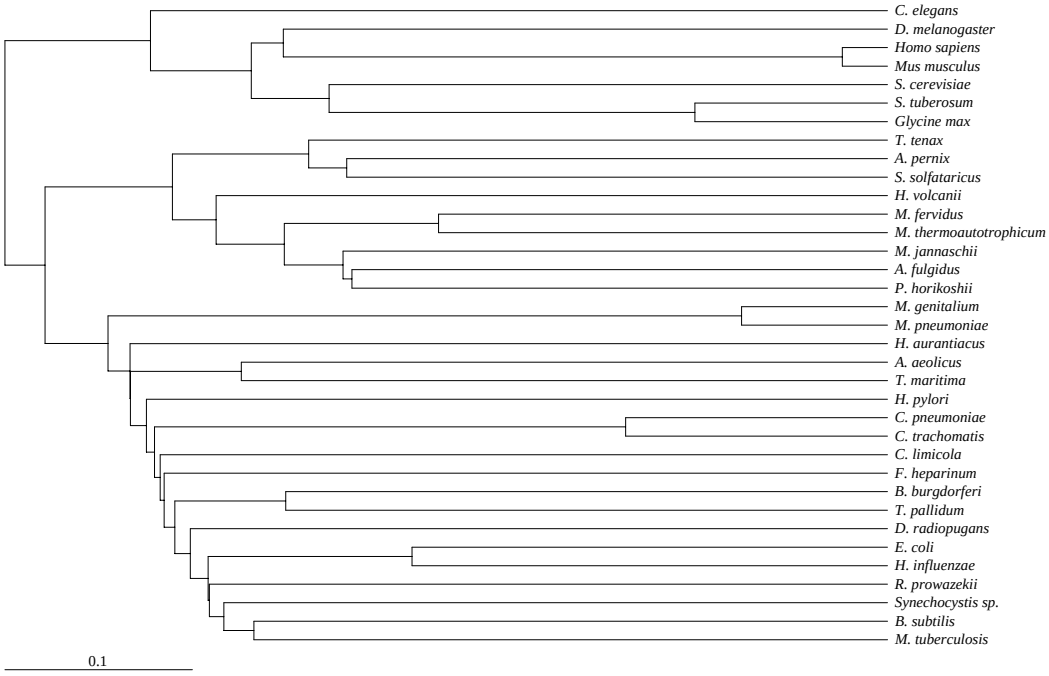
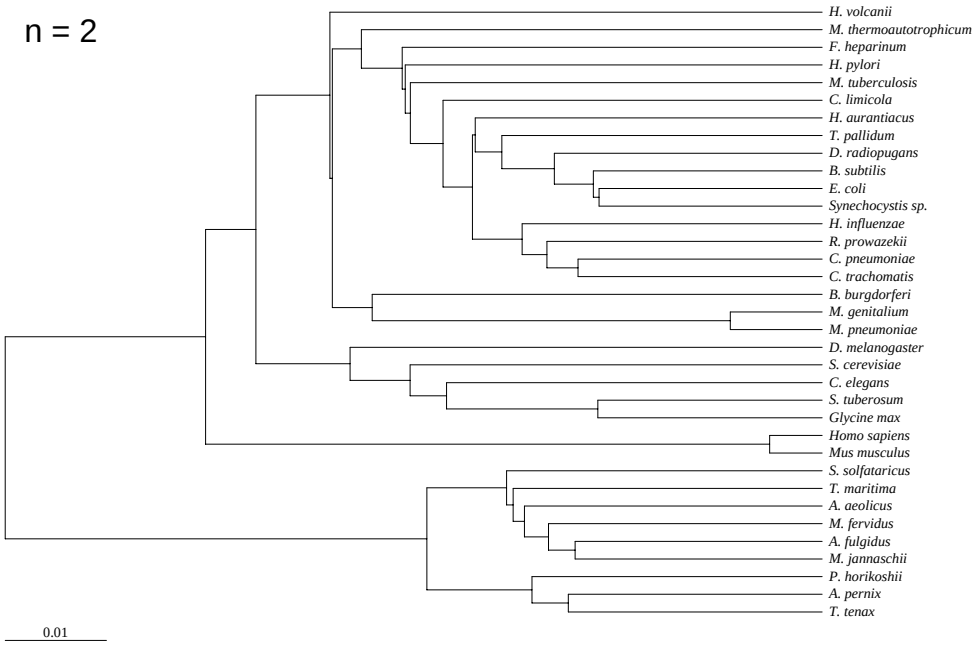


Figure 3: 35-organism tree constructed using the UPGMA method based on the 8-distance.



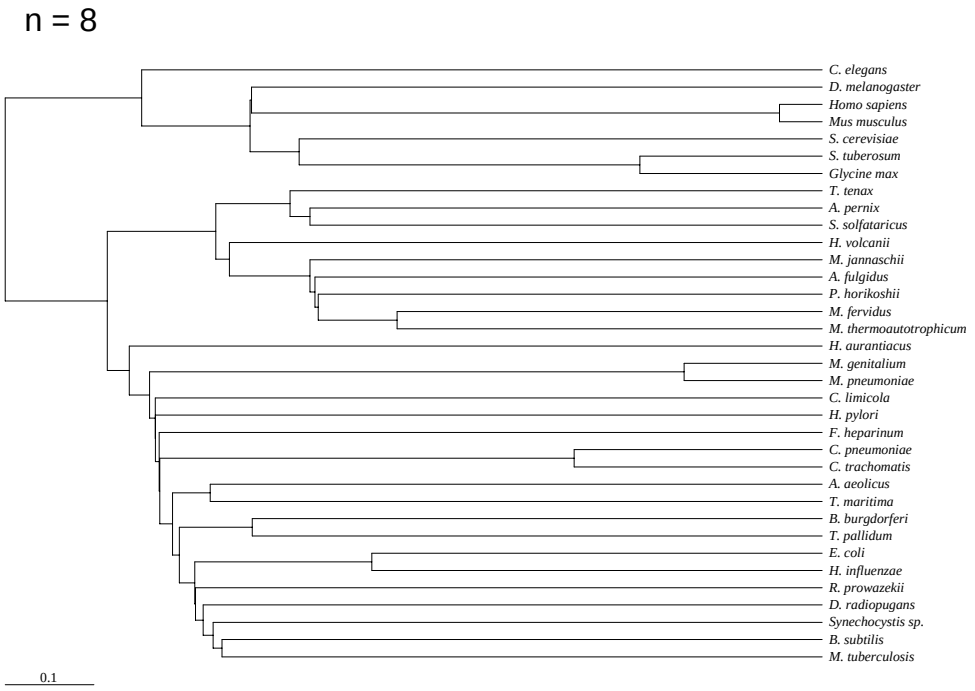


Figure 4: 35-organism trees constructed using the FC method based on the 2- and 8-distances.

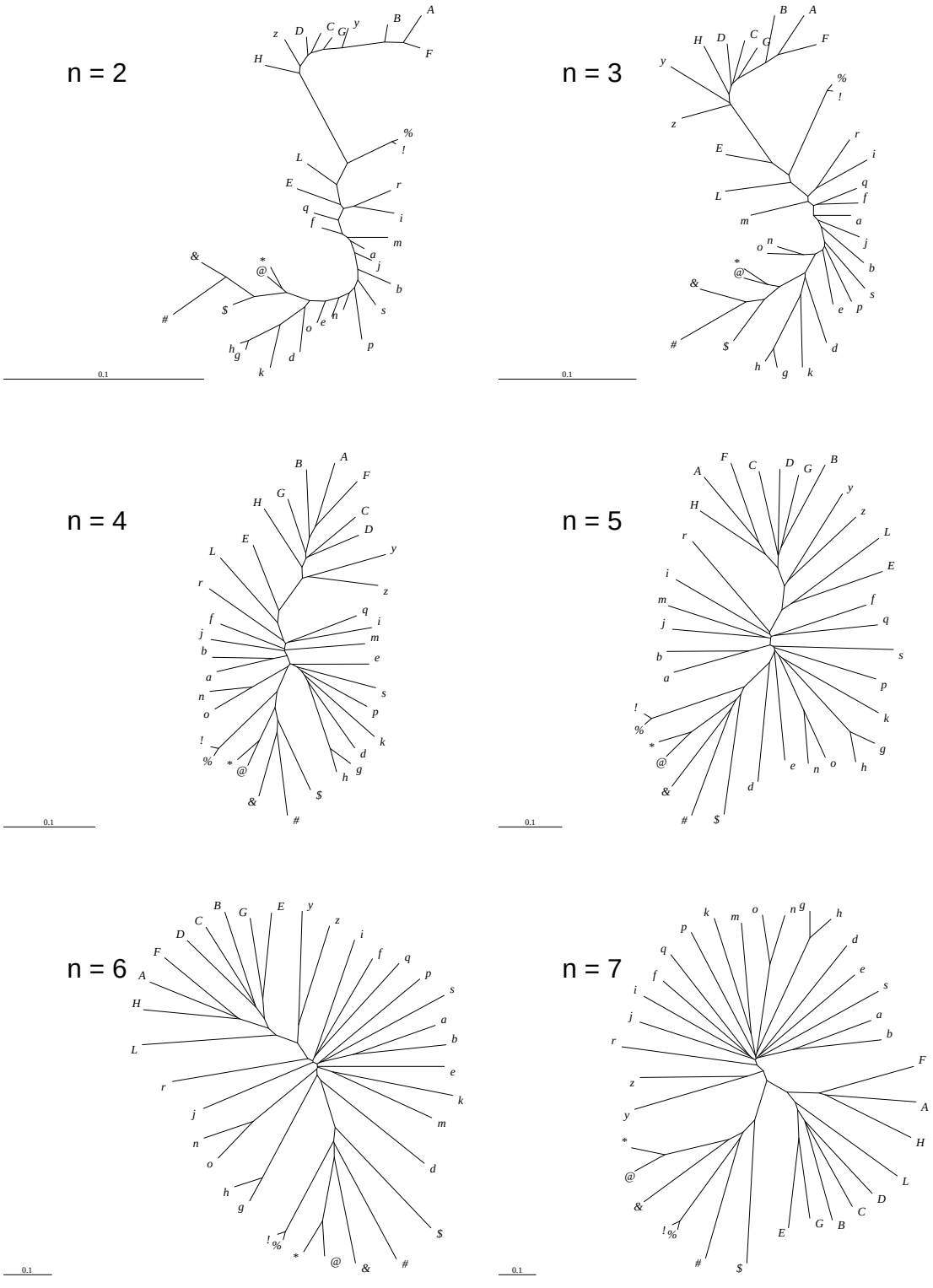


Figure 5: Unrooted 35-organism trees constructed using the NJ method based on n-distances, n=2 to 7. Code for organisms is given in Table 1: upper-case Roman alphabets for archaeons, lower-case alphabets for bacteria, non-alphabet symbols for eukaryotes. y and z stand for the two thermotogales.

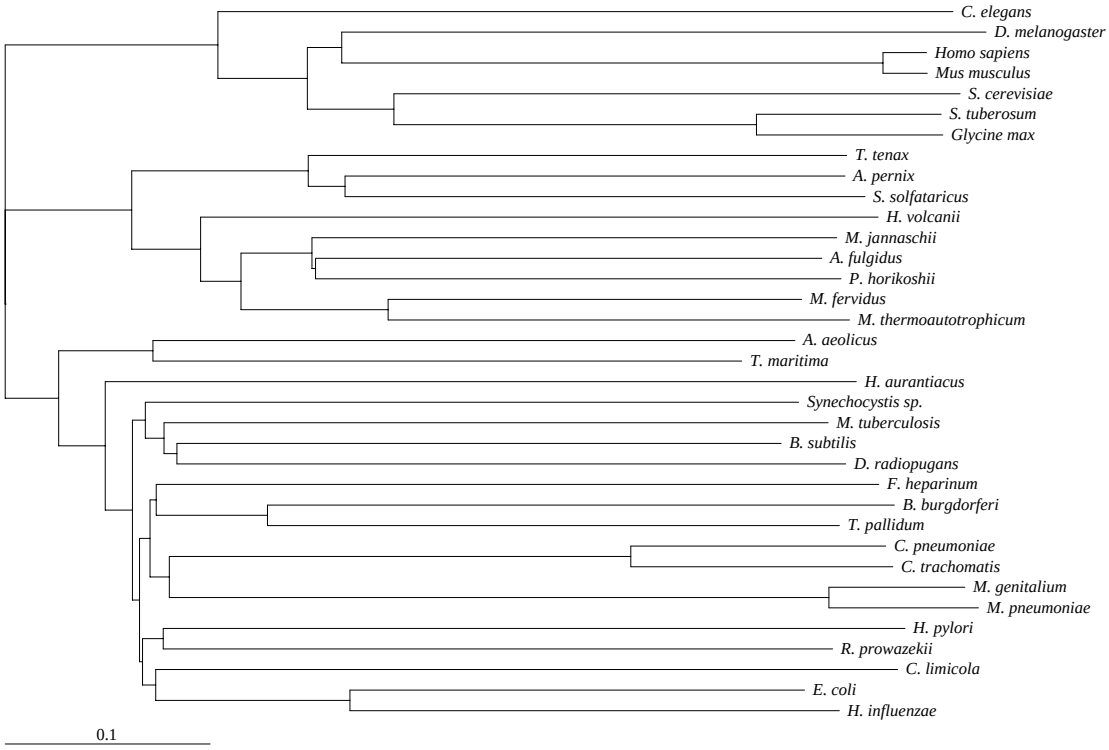


Figure 6: 35-organism tree constructed using the NJ method based on the 7-distance with Eukarya as the outgroup.