Biomedical Sciences Today
An open access peer reviewed journal
MDT Canada press
http://www.mdtcanada.ca/bmst.html

Regular Article

Evolutionary Tree Based on Oligonucleotide Frequencies and Conserved Words in 16S and 18S Ribosomal RNA

Li-Ching Hsieh[1], Chih-Yuan Tseng[2], Liaofu Luo[3]*, Mingwen Jia[3], Fengmin Ji[4], Hoong-Chien Lee[5,6*]
[1]Institute of Genomics and Bioinformatics, National Chung Hsing University, Taichung City, Taiwan 402
[2]MDT Canada Inc., Edmonton, AB Canada T5B 2P2
[3]Department of Physics, Inner Mongolia University, Hohhot, 010021, China
[4]School of Life Science, Northern Jiaotong University, Beijing, China
[5]Department of Physics, Chung Yuan Christian University, Zhongli, Taiwan 32023
[6]Department of Biomedical Science and Engineering, National Central University, Chungli, Taiwan 32001
*Correspondence Email: lfluo@nmg2.imu.edu.cn (Liaofu Luo); hclee12345@gmail.com (HC Lee)

Received 12 November 2015; Revised 1 December 2015; Accepted 1 December 2015; Published 25 December 2015

Editor: Mohammad Ashrafuzzaman

Abstract

Sequence distances are defined in terms of the differences in the oligonucleotide frequencies of length *n*. Such *n*-distances are used to construct phylogenetic trees from a set of thirty-five 16S (18S) rRNA sequences. The quality of the trees generally improves with increasing *n* and reaches a plateau at *n*=7 or 8. The best *n*-distance trees are compatible to trees based on sequence alignment, suggesting that highly overrepresented 7-mers and 8-mers are closely related to rRNA evolution. Out of the 4^7=16384 7-mers, 612 are identified as those whose relative frequencies correlate strongly with the 35×35 *n*-distance matrix. These evolution-related 7-mers are used to identify "conservative words", oligonucleotides whose frequencies and loci are common to at least 85% of organisms preselected to represent a domain. The structural meaning of some of these conservative words is discussed.

Keywords: Oligonucleotide frequency, ribosomal RNA sequence, sequence distance, phylogenetic tree, conservative words, evolution

Biomedical Sciences Today
An open access peer reviewed journal
MDT Canada press
http://www.mdtcanada.ca/bmst.html

1. Introduction

The investigation of oligonucleotide correlation in DNA sequences plays an important role in understanding the genetic language [1,2]. When a sequence has an oligonucleotide with a frequency much higher or much lower than expected in a random sequence, this oligonucleotide is referred to as a preferred or forbidden "word" in the sequence. Many dinucleotide and trinucleotide preferred words were found and their evolutionary meaning have been discussed [3-5]. Forbidden and preferred words of nucleotides six to nine chemical letter long in some genomes have also been studied [6,7]. DNA uptake signal sequences in a number of human pathogens, typically nine to ten letters long, have been studied in great detail [8-10]. Compilations of biologically meaningful words of from the statistical analysis of oligonucleotide frequencies in DNA sequences have been made [11-13]. In what follows we call an n-letter word an n-mer.

Preferred words with biological functions are expected to be evolutionarily conservative, which may be detectable in a sequence-based reconstruction of a phylogenetic tree. Several questions regarding this aspect can be addressed immediately. (1) Are oligonucleotide frequencies useful for studying evolution? (2) Specifically, can they be used to construct a good phylogenetic tree? (3) Are there preferred or forbidden oligonucleotides that have played special roles in evolution? (4) If so, how can they be identified? Here we answer the first three questions in the affirmative and for last question we propose a method to find evolution-related conserved words.

Multiple sequence alignments is a standard method widely used to study sequence similarity and to search for conservative words. This approach is extremely powerful yet not geared to search for fully conserved words located at nearly the same sites in a large set of sequences. Another shortcoming of this method is that the computation time required grows exponentially with increasing sequence length, the size of the sequence set, and decreasing similarity. Here we explore another approach based on oligonucleotide frequencies. A new type of sequence distance, the *n*-distance, is devised. It measures the difference in the frequency distributions of n-mers between two sequences. Computational time for the n-distance between all pairs of sequences in a set of sequences grows as fixed powers of sequence length and of the number of sequences.

A set of thirty-five 16S rRNA (for archaeons and bacteria) and 18S rRNA (for eukaryotes) sequences is used to investigate properties of the *n*-distance. The clock-like property in the evolution history of these sequences makes the sequence suitable objects for searching for evolution relations among organisms [14]. The universal phylogenetic tree based on rRNA is generally recognized as a valid representation of organismal genealogy [15]. We use the *n*-distance to construct phylogenetic trees, called n-trees, for the 35 organism and compare them with alignment(-based) trees. We show that n-trees have qualities similar to those of alignment-trees when n≥7.

An important feature of the phylogenetic tree is its set of deepest branchings, which indicates the earliest major diversions of domains and kingdoms. These branchings occurred in an era when rRNA and a translation apparatus more primitive than what they are today were probably the main biomolecular machineries. If there still exist some functional sites in present-day rRNAs that have been fully conserved throughout the long history of rRNA, then we expect their conservation patterns across species to be correlated with the deepest branchings. Because random mutations are the major driving force of sequence evolution such sites, or words, if they do exist, will lie deeply hidden in a huge background. We devise an algorithm for identifying these words and use it to find many conserved words. A number of these words are fully conserved in a very large number of organisms, to such an extent that they may be said to characterize a whole domain. Possible biological meanings of some of these words are explored through inspection of the secondary and tertiary structures of ribosomal RNAs.

2. Methods

2.1 Database

The first part of our algorithm to find conserved words is to use *n*-distance to construct a phylogenetic tree that is a representation of the Tree of Life. We choose thirty-five organisms - 9 archaeons, 19 bacteria and 7 eukaryotes - for this purpose. The organisms are chosen from among those whose genomes have been completely sequenced such that a large number of classes of organisms is covered. For representation of important subclasses that are still missing, some organisms whose genomes are not completely sequenced are then chosen. All sequence data are taken from the GenBank [16]. The selected

Biomedical Sciences Today
An open access peer reviewed journal
MDT Canada press
http://www.mdtcanada.ca/bmst.html

organisms and the GenBank accession number of their 16S or 18S rRNA sequences are listed in Table 1. In the table archaeons are coded by upper-case Roman alphabets, bacteria by lower-case alphabets and eukaryotes by non-alphabet symbols. For the culling of conserved words the 16S rRNA sequences of an additional set of sixty-one prokaryotes - 20 archaeons and 41 bacteria - are selected from the prokaryotic tree of [17]. These are given in Table 2.

Table 1. The 35 organisms, their single-letter or 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 M61706 M81326
q	*Deinococcus radiopugnans*	Y11334
r	*Herpetosiphon aurantiacus*	M34117
s	*Chlorobium limicola*	Y08102
y	*Aquifex aeolicus*	AE000657
z	*Thermotoga maritima*	AF001701
%	*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

The benchmark against which the quality of n-trees are tested is the “life-tree”' extracted from the consensus Tree of Life by removing all organisms except the 35 organisms included in Table 1. The three-domain topology of the Tree of Life is from [14]. Its Archaea and Bacteria branches are reconstructed from the prokaryotic tree of [17] and its Eukarya branch is from [18] and [19]. However, because the property of a dendrogram depends significantly on the size of its entries, an “alignment-tree” based on multiple alignment of the 35 rRNA sequences is also used as control and for direct comparison with the n-trees. This tree is constructed through the software package OMIGA 1.13 [20] where default parameters are used.

Table 2. Extended set of 61 prokaryotes including 20 archaeons 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. [illegible]*	*G. [illegible]*	*F. nodosum*
T. melanesiensis	*T. roseum*	*C. aurantiacus*	*T. [illegible]*
T. thermophilus	*D. radiodurans*	*P. hollandica*	*A. cylindrica*
N. sp.	*L. monocytogenes*	*K. zopfii*	*G. haemolysans*
M. hyopneumoniae	*M. [illegible]*	*M. fermentans*	*M. gallisepticum*
N. [illegible]	*M. [illegible]*	*F. aquatile*	*F. columnare*
F. [illegible]	*C. vibrioforme*	*T. pallidum*	*F. [illegible]*
F. litoralis	*S. aurantia*	*C. psittaci*	*P. [illegible]*
I. pallida	*C. jejuni*	*W. succinogenes*	*R. rickettsii*
E. [illegible]	*W. [illegible]*	*V. parahaemolyticus*	*P. vulgaris*
E. carotovora			

2.2 The *n*-distance

Denote the probability of letter *a* (*a*=A, G, C or T) occurring in a sequence by p_a, and the joint probability of letters *a* and *b* occurring sequentially in the sequence by p_{ab}. In general, if $\sigma = abc\Lambda$ is an n-mer, denote joint probabilities of the letters in σ, or relative frequency of σ, occurring in the sequence by p_σ. We assume all sequences to be circular in the calculation of joint probabilities. For given n the probabilities satisfy the sum-rule $\sum_\sigma p_\sigma = 1$, where the summation is computed over the set $\{\sigma\}_n$ of all the 4^n n-mers. So long as 4^n is less than the sequence length *N*, the set $\{\sigma\}_n$ with increasing n is an increasingly fine-grained characterization of a sequence. Given two sequences Σ and Σ' with joint-probability sets $\{p_\sigma\}$ and $\{p'_\sigma\}$, respectively, the quantity

$$D^{(n)}(\Sigma, \Sigma') = \sum_{\sigma \in \{\sigma\}_n} \left| p_\sigma - p'_\sigma \right| \qquad (1)$$

is defined as the n-distance between two sequences. The computation of an *n*-distance does not involve any sequence alignment. For a set of sequences $\Sigma_\alpha, \Sigma_\beta, \Lambda$, we define an *n*-distance matrix $D^{(n)}$ whose matrix elements are

Biomedical Sciences Today
An open access peer reviewed journal
MDT Canada press
http://www.mdtcanada.ca/bmst.html

$D_{\alpha\beta}^{(n)} = D^{(n)}(\Sigma_\alpha, \Sigma_\beta)$. An \$n\$-distance is well defined for sequences that are of different lengths and are not aligned. The value of $D_{\alpha\beta}^{(n)}$ ranges from 0, when the two sequences are identical, to 2, when they are totally dissimilar.

2.3 The *n*-tree

For $2 \le n \le 9$, \$n\$-distance matrices for the 35 organisms in Table 1 are computed. Dendrograms, or \$n\$-trees, are then constructed from the distance matrices using the neighbor-joining (NJ) method, the unweighed pair-group mean arithmetic (UPGMA) method [21] and the fuzzy clustering (FC) method [22]. The software package PHYLIP version 3.5c was used for tree construction and plotting [23]. The third method, 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 as follows.

Given a distance matrix *D* (representing any $D^{(n)}$) we define a similarity matrix $S = 1 - D/2$. Because *D* is symmetric with vanishing diagonal elements, *S* is symmetric and reflexive (unitary diagonal elements). An element of *S* measures the closeness of the two sequences, and has value 0 for two total dissimilar sequence and value 1 for identical sequences. From *S* we compute a fuzzy similarity matrix $\hat{S}$ and then from $\hat{S}$ use the method of fuzzy clustering to compute a fuzzy equivalence matrix $\hat{E}$. From $\hat{E}$ we then use the alpha-cut technique to construct a partition tree.
Details are given in [22].

2.4 Evolution-related 7-mers

We consider the (35×34/2=) 595 different elements of a distance matrix *D* to be independent. Let $D_{sw}(\sigma_0)$, a single-word distance matrix associated with the \$n\$-mer σ_0, be the matrix whose elements are defined as in Eq. (1), except the summation on right-hand-side of the equation takes only a single term from $\sigma = \sigma_0$. We define a correlation coefficient between *D* and $D_{sw}(\sigma_0)$ as

$$Cor(\sigma_0) = Cov(D, D_{sw}(\sigma_0))/(Var(D)Var(D_{sw}(\sigma_0)))^{1/2} \quad (2)$$

For a sampling size of 595, a value for *Cor* that is greater than the threshold value (at 99% confidence level) of 0.11 is expected to play a significant role in the evolution process. For this work, we designate those \$n\$-mers whose correlation coefficients are larger than a cut-off value of 0.30 as evolution-related \$n\$-mers (EROs). In a first round of search we consider only EROs for \$n\$=7 (ER7s) and identified 612 of these.

2.5 Conserved words in the three domains

Conserved words (CWs) in the three domains are identified using the previously determined set of ER7s through following three steps: 1) Find EROs with \$n\$>7 but limiting the search to those that has an ER7 as a subsequence; 2) Identify an ERO as a candidate CW when its sites are the same on at least two rRNAs among those given in Table 1; 3) A candidate CW is designate a CW if it appears at approximately the same sites in a majority of rRNAs among all the prokaryotes listed in Tables 1 and 2. These steps are described in detail below.

(1) *Search for all EROs*. We use Eq.(2) to search for EROs for $8 \le n \le 13$, limiting the search to words that has at least one ER7 as a subsequence.

(2) Search for candidate CWs. We take an ERO as query and match it against the rRNA sequences in Table 1 and 2 using the BLAST program [24]. If the query occurs at nearly the same sites - difference less than 100 bases - in at least two organisms, it is designated a candidate CW. If an ERO only has a common site (or common sites) that is (are) also the common site(s) of a longer ERO(s), then the shorter ERO is not designated a candidate CW. For example, the 13-base ERO GCGGTGAATACGT and its subsequence the 8-base ERO GCGGTGAA are used as queries against the subjects *D. radiopugans* and *F. heparinum* and in BLAST searches and the following results are obtained:

Query	Length	Sites of query in *D. radiopugans*	Sites of query in *F. heparinum*
GCGGTGAATACGT	13	1316-1328	1360-1372
GCGGTGAA	8	1316-1323	1360-1367
		–	683-690

Sites 1316-1328 and 1360-1372 are approximately the same so the 13-mer GCGGTGAATACGT is designated a candidate CW (for *D. radiopugans* and *F.*

Biomedical Sciences Today
An open access peer reviewed journal
MDT Canada press
http://www.mdtcanada.ca/bmst.html

heparinum). The 8-mer GCGGTGAA occurs in *F. heparinum* at two sites and occurs in *D. radiopugans* at one site. The site 1360-1367 in *F. heparinum* is common to the site 1316-1323 in *D. radiopugans* whereas the site 683-690 in *F. heparinum* is a singleton (as far as the two organisms under discussion are concerned). Since the common site is embedded within the common site of the 13-mer, we designate the 13-mer a candidate CW but not the 8-mer. BLAST gives matches that are not identical to the query; some matches contain mismatched sites, or gaps, or both. We limit the possible number of candidate CWs by discarding such non-perfect matches.

(3) *Identify CWs*. If a candidate CW is conserved in more than 85% of the rRNA sequences of the same-domain organisms, listed in Table 1 and 2, then it is identified as a CW. Because Table 2 contains only archaeons and bacteria, the CWs are said to be archaean or bacterial, or both, as the case may be, but never as an eukaryotic CW.

3. Results

3.1 The life-tree and alignment-tree

The early divergence of the Tree of Life is said to be problematic [25-27], yet it seems to be settling on a consensus branching pattern (Bacteria, (Archaea, Eukarya)) for the three domains [28,29]. Part of the reason for this discrepancy is the following. Because 18S rRNAs in the eukaryotes, being about 1800 bases long, are about 250 bases longer than the average length of prokaryote 16S rRNAs, long gaps in the consensus 16S rRNA sequence necessarily appear when an 18S rRNA is aligned against a 16S rRNA. Unless such gaps are masked after the alignment and ignored in the computation of similarity score, the score between sequences lying across the Eukarya-Prokarya divide will be markedly less than that between sequences within Eukarya or Prokarya. The result is that not masking reduces the Bacteria-Archaea distance relative to the other two interdomain distances. In this work the Tree of Life refers to the tree constructed by [19] based on sequence alignment with masking; the 35-organism life-tree shown in Fig. 1 is obtained from the Tree of Life by pruning all branches not leading to one of the 35 organisms; the alignment-tree shown in Fig. 2 is obtained from the 35 rRNA sequences using sequence alignment without masking.

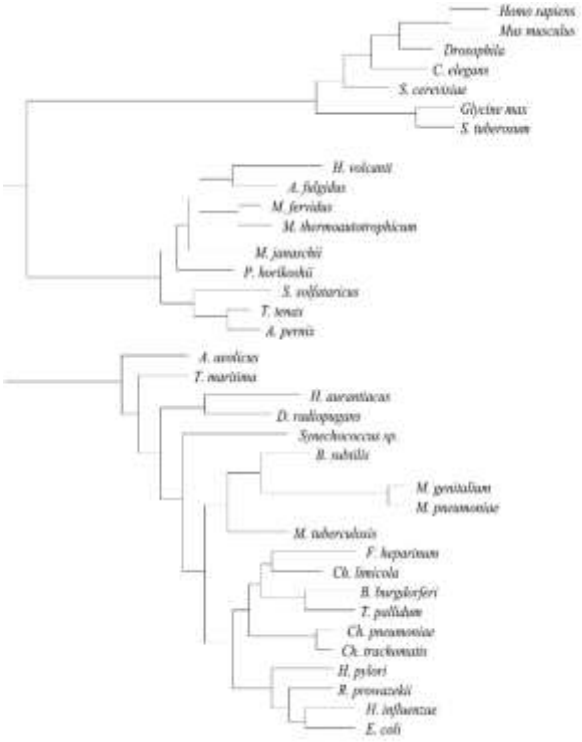


Figure 1. The life-tree, obtained by removing from the Tree of Life all except the 35 organisms listed in Table 1. The Archaea and Bacteria domains are reconstructed from [17], and the Eukarya domain is from [18; 19]. Branch lengths are only approximately to scale.

Figure 2. The alignment-tree, obtained from alignment of 16S/18S rRNA sequences of 35 organisms listed in Table 1 without masking and with Eukarya as out-group.

Biomedical Sciences Today
An open access peer reviewed journal
MDT Canada press
http://www.mdtcanada.ca/bmst.html

3.2 The n-trees

Fig. 3 shows the unrooted trees constructed through the NJ method from *n*-distances with n= 2 to 7 (henceforth call the n-trees). The quality of the trees improves with increasing n when n<7, and reaches a plateau around n=8 (not shown). This trend is general and does not depend on methods of tree construction. The n=7 and 8 trees are found to be the best in the sense that they are most similar to the alignment-tree shown in Fig. 4. The overall pattern is: recognizable Archaea from n=2, formation of Eukarya as a separate group from n=4, and formation of the three domains from n=7. The n=7 tree is shown in detail in Fig. 4. Interestingly, the three domains are recognizable when tree are constructed via the FC method even on the 2-tree (top tree in Fig. 5). The best trees constructed using the FC (bottom tree in Fig. 5) and UPGMA (Fig. 6) methods occur at n=8. The main branchings on these trees are similar to those on the best NJ tree (Fig. 4), but the topology for the Bacteria domain on the NJ tree is closest to the life-tree. The comparison of these best trees with the life-tree (Fig. 1) is summarized in Table 3, where the number of moves needed to bring the branching pattern into agreement with the life-tree is given in the last column.

Table 3. Comparison of first few levels of branchings of the three domains 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 Tree of Life.

Tree	Branching pattern		Moves	
	Eukarya	Archaea		
Tree of Life	((@*)(&(\$(#(%!)))))	((H(AF))(B(D((CL)(EG)))))	-	-
Alignment-tree	((\$((#(%!))((@*)&))))	((H(AF))(L((D(BC))(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(qj)((fl(gh))i))(((no)(km)(ps))((ab)ed)))))))		-	
Alignment-tree	((yz)((q(fl))(ij(no)))(r((km)((ps)((gh)((ab)ed))))))))		3	
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(fl)e(ab))))))))		4	
n=8 UPGMA tree	((gh)(r((yz)(d((no)(sp(km)qej(fl)(ab)))))))		5	

3.3 Evolution-related 7-mers

Six-hundred and twelve evolution-related 7-mers (ER7s) are found; these are listed at the website http://pooh.phy.ncu.edu.tw/~hclee/ER7/. The list includes 311 with $0.30 \le Cor \le 0.4$, 175 with $0.40 \le Cor \le 0.5$, and 126 with $Cor \ge 0.5$. The occurrence frequencies of sixty ER7s with $Cor \ge 0.52$, grouped into 19 sets according to their frequency pattern and in descending order in *Cor* value, are shown against the 35 organisms in Table 4. All n-mers all given with the 5' end on the left. When a set contains more than one ER7, all the ER7s share the same frequency pattern. For example, twenty-eight ER7s in set number 4 have the same frequency pattern: each occurs twice in the eukaryotic rRNAs and does not occur in prokaryotic rRNAs. We classify the ER7s into six categories according to the domain or domains to which the organisms containing a common ER7 belong: I - Archaea and Eukarya; II - Eukarya; III - Archaea and Bacteria; IV - Bacteria; V - Archaea; VI - Eukarya and Bacteria. The general trend is that the *Cor* values decreases with increasing category number. Six of the sixty ER7s in Table 4, including the four ER7s with the greatest *Cor* values ($Cor \ge 0.59$), that is all the ER7s in first three sets in Table 4, belong category I. Thirty-six of the ER7s in Table 4, including the 28 ER7s in set 4, and all having $Cor \ge 0.52$, belong to category II. In comparison all ER7s in the last two categories, V and VI, have *Cor*<$0.4 (see website cited earlier).

3.4 Conservative words

The number of EROs having at least an ER7 as a subsequences are too numerous to be listed here. EROs that are located at nearly the same sites - deviation must be less than 100 bases - on a large number of 16S (or 18S) rRNA sequences are designated conserved words (CWs). Information pertaining to the nine most widely conserved CWs in Bacteria and Archaea are given in Table 5 [30]. The two most widely conserved CWs are the 8-mer AACGAGCG and 13-mer GGATTAGATACCC. AACGAGCG is in the rRNAs of all 89 prokaryotes listed in Tables 1 and 2 except the archaeon *T. acidophilum* and the bacterium *C. vibrioforme*; it is also found in the eukaryote *C. elegans*. GGATTAGATACCC is in all the 89 prokaryotes except the archaeon *M. igneus* and is absent in the eukaryotes. Although several of the CWs appear in all or most of the eukaryotes listed in Table 1, they are not

Biomedical Sciences Today
An open access peer reviewed journal
MDT Canada press
http://www.mdtcanada.ca/bmst.html

designated eukaryotic CWs because only seven eukaryotes are included in this study.

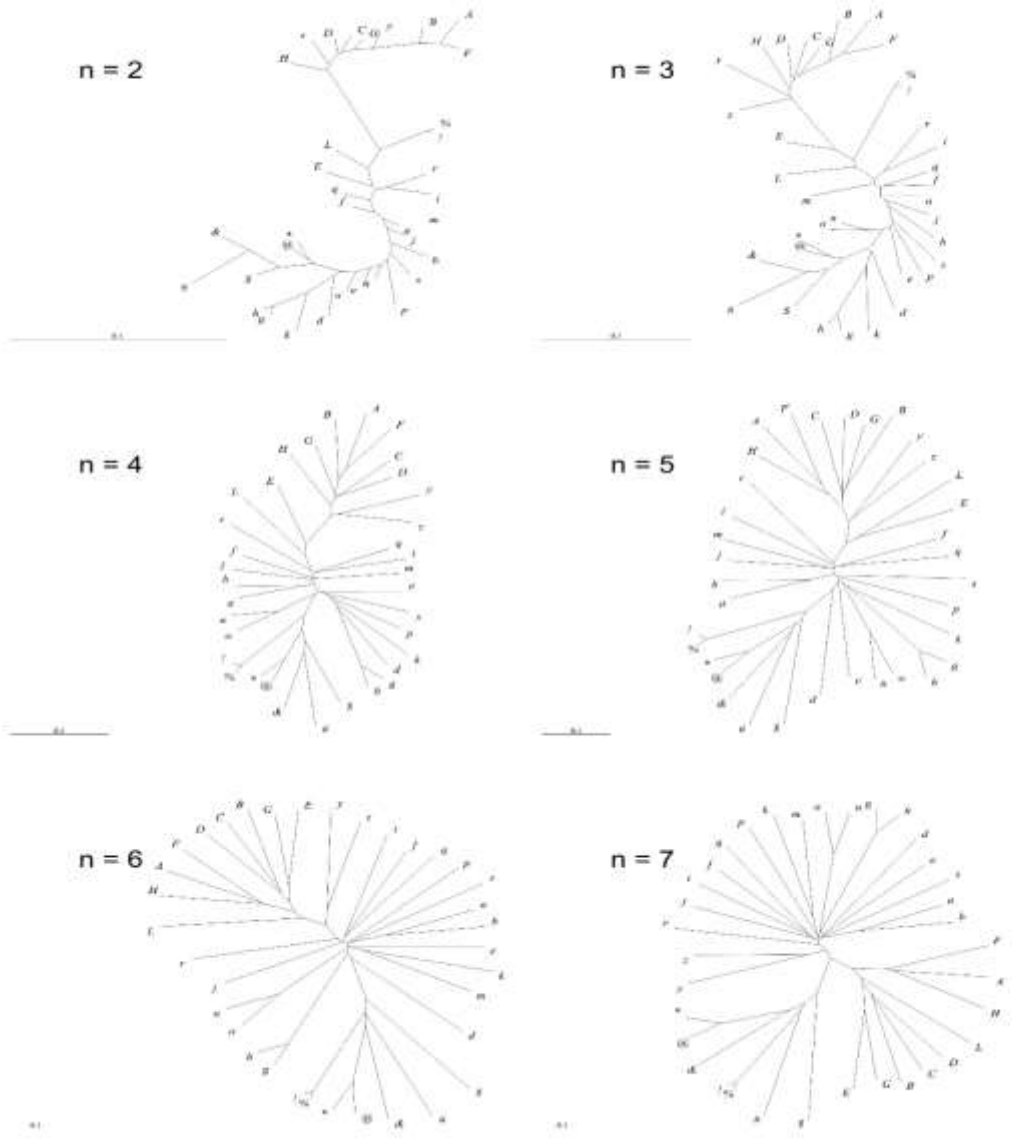


Figure 3. Unrooted 35-organism *n*-trees constructed using the NJ method based on *n*-distances, *n*=2 to 7. Code for organisms is given in Table 1: upper-case Roman alphabets for archaeons, lower-case alphabets for bacteria, non-alphabet symbols for eukaryotes; *y* and *z* are the two thermotogales mentioned in the text.

Biomedical Sciences Today
An open access peer reviewed journal
MDT Canada press
http://www.mdtcanada.ca/bmst.html

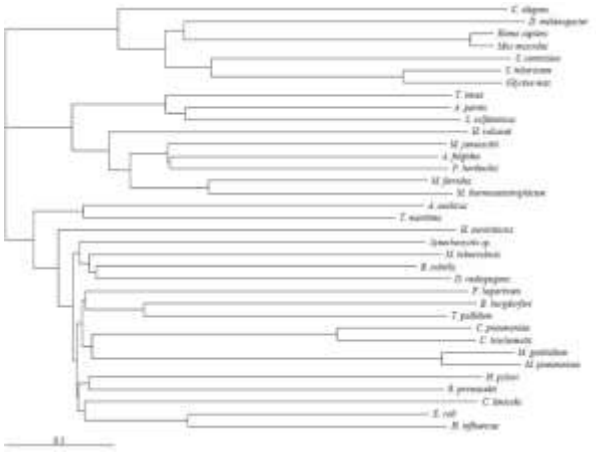


Figure 4. 35-organism (*n*=7)-tree constructed using the NJ method with Eukarya as outgroup.

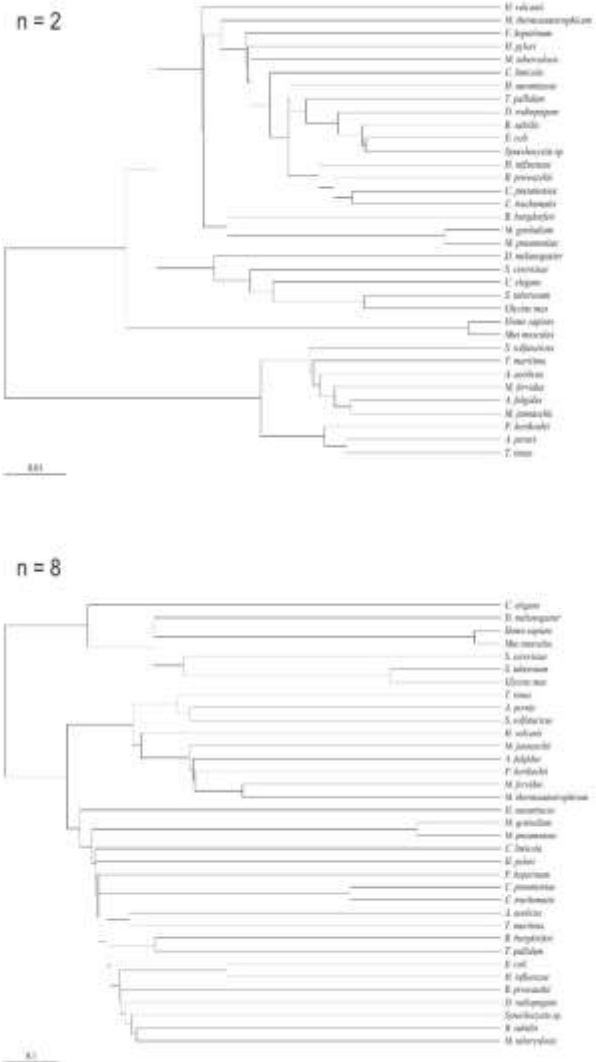


Figure 5. 35-organism *n*-trees for *n*=2 and 8 constructed using the fuzzy clustering method.

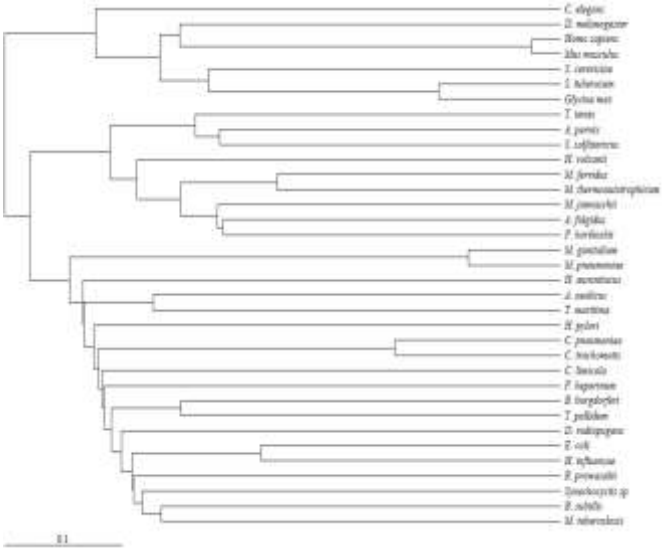
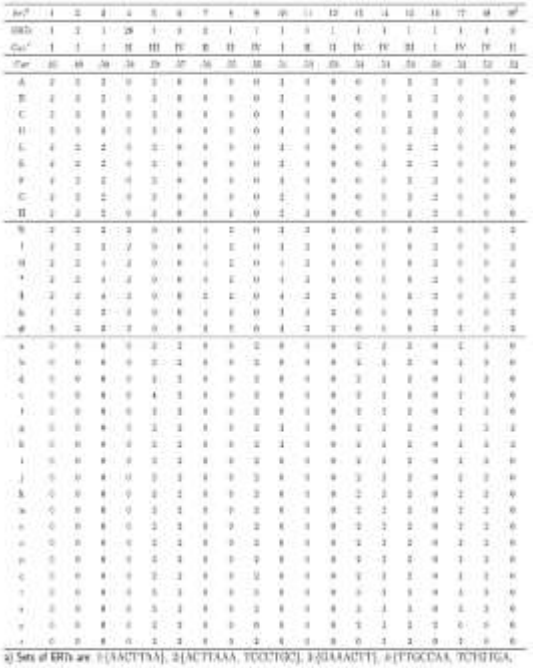


Figure 6. 35-organism (*n*=8)-tree constructed using the UPGMA method with Eukarya as outgroup.

Table 4 Evolution-related 7-mers with the largest correlation coefficients and their frequencies (normalized to 3000 bases) in the 35 organisms of Table 1.



a) Sets of ERTs are I:{AACTTAA}, 2:{ACTTAAA, TCCTTGC}, 3:{GAAAGTT}, 4:{TTGCCAA, TCTGTGA, TTTGTCT, TCAAGAA, TGATGCC, TAATCAA, CTTCTTA, CTCAACA, CTGTGAT, CTGGTGC, CATGCAT, GTGTGTG, GGTCTGT, GGCCGTT, GATTTGT, TTTGTAC, ATGGTAG, ATATGCT, ATAACAG, TTCCACC, AGTATCG, AGAAAGG, AATCAAG, AATGAGT, TTCATTA, AACAGCG, AACCATT, AAAGCAT}, 5:{CTACTGG}, 6:{CGCAACC, GCGCAAC, CCGCAAC}, 7:{GGTAATT, GTAATTC}, 8:{GTATGGT}, 9:{GCACAAG}, 10:{AAACTTA}, 11:{CGGAGAG}, 12:{TAATTCC}, 13:{AATTCGA}, 14:{TCCCGCA}, 15:{TATTAGA}, 16:{CAGCCGC}, 17:{GTTCTAC}, 18:{ACTCCTA, TAATTCC, TCCTACG, TGTCCTC}, 19:{TTTGACG, TCAAAAA, AGGGTTC}. b) Five *Cor*=0.52 ERTs (category in brackets) not listed because of lack of space are: GACCCCA (IV), CGCAACC (IV), CCGCACC (II), GCCACCA (II) and CGAGCCT (I). c) Category, see text for definition.

Biomedical Sciences Today
An open access peer reviewed journal
MDT Canada press
http://www.mdtcanada.ca/bmst.html

Table 5. The nine most widely conservative words[a]

No	CW	Site	Domain	Remarks [c,f]
1	GGATTAGATACCC	785-797 [e]	Arch./Bact.	Arch.: not in *M. igneus*. On end loop near H24, cross link between A(794) and H23(693-696).
2	AACGAGCG	1102-1109 [e]	Arch./Bact.	Arch.: not in *T. acidophilum*; Bact.: not in *C. tuberiformis*; also found in *C. elegans*. On H35.
3	GACGGTGAC	711-719 [d]	Arch.	Not in *H. halobium*.
4	CCTTGCACACAC	1352-1363 [d]	Arch.	Arch.: not in *M. mazeiensis*; Bact.: only in *A. pyrophilus* and *A. aeolicus*.
5	AAACTCAAA	907-915 [e]	Bact.	Arch.: only in *D. mobilis*; Bact.: not in *M. hypermanniae*, *M. soehngenii* and *M. hexonica*. Between H27 and H2.
6	TGGGTTAA	1086-1093 [e]	Bact.	Not in *T. pallida*, *P. staleyi*, *E. hirota*, *F. columnare*, *F. aquatile*, *M. hypermanniae*, *M. sodalis*, *M. hexonica* and *C. aurantiaca*. On H37 and its downstream end loop; crosslink between H(1086-1094) and H(1161-1168).
7	ACCACCAG	674-681 [d]	Arch.	Arch.: all of Crenarchaeota; Bact.: only in *C. aurantiaca*. Also in Eubs. except *C. elegans* and *S. cerevisiae*.
8	GTAGTCCCG	759-767 [d]	Arch./Bact.	Arch.: all of Crenarchaeota; Bact.: only in *A. pyrophilus* and *A. aeolicus*.
9	CCCGTCGC	1366-1373 [d]	Arch.	Arch.: all of Crenarchaeota; also found in Eubs.

a) The conservative words listed are conservative only in Bacteria and/or in Archaea. They are not found in Eukarya near the sites quoted in the table unless otherwise indicated in the Remarks. Organisms referred to are restricted to the 96 species included in Tables 1 and 2. b) Structural information refers to 16S rRNA of *E. coli* (32). c) Sites given refer to those in *E. coli*. Sites of conservative words in organisms other than *E. coli* are near the sites quoted. d) Sites given refer to those in *T. ferrus*. Sites of conservative words in organisms other than *T. ferrus* are near the sites quoted.

4. Discussion

4.1 The best n-trees are as good as the alignment-tree

The three best n-trees (N=7, 8, 9) all separate the three domains cleanly and agree in their general features with the life-tree and the alignment-tree. Yet detail branching patterns on all the trees differ; such patterns are known to be extremely sensitive to small changes in the distance matrix and to tree construction methods. When finer details are considered Table 3 shows the n=7 NJ tree to be the best n-tree.

If two sequences have a high degree of similarity, their n-distances should be small for every n. This effect is shown clearly in 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. Irrespective of the method used for tree construction, these pairs are the closest neighbors on every tree with n>2. Four other slightly less 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*, whose aligned sequences are 89%, 86%, 80% and 77% identical, respectively, are closest neighbors on the best n-trees. On the best n-trees the archaeons correctly divide into a group of three crenarchaeotes and a second group of six euryarchaeotes. However, on the n-trees the positions of *H. volcanii* and *P. horikoshii* are inverted relative to the life-tree. On the n-trees the eukaryotes form a group by themselves for good starting at n=4. Eukaryotes take their final branching pattern at n=7 or 8. The branching pattern on the best trees is identical to that on the alignment-tree: (nematode((fly(mammals))(yeast(plants)))), compared to that on the life-tree (plants(yeast(animals))) [19,27]. Thus for the set of test eukaryotes the 7- and 8-distances are as good as distances determined by sequence alignment without masking.

Previous studies have designated the thermotogales *A. aeolicus* and *T. maritima* to be among the deepest branching bacteria [31,32]. With the complete sequencing of the genomes the lineage of the two organisms are under re-examination. Recent analyses indicate that both thermotogales share a common ancestor with Bacteria for a majority of genes involved with housekeeping functions such as transcription, translation, DNA replication and cell division. In addition, they also inherited about half of their genes involved with metabolic functions from the ancestor of Archaea [33,34]. The mixed heritage of the thermotogales has been taken to be evidence of extensive horizontal gene transfers between Archaea and Bacteria. The ambiguous nature of the lineage of the thermotogales is evident on our constructed trees even though the trees are based on a single gene, the 16S/18S rRNA. On all three types of constructed trees they are placed in Archaea when $n\le6$, often as an out-group, and go over to Bacteria when n ≥ 7. This suggests that traces of the ambiguity is distributed over the genomes in unit that are smaller than the genes.

We can now answer questions (1) and (2) raised in the introduction: Frequencies of occurrence of n-mers and joint probabilities of n-mers in the form of n-distances are useful for studying sequence evolution and can be used to construct fair quality phylogentic trees. For n=7, there are 4^7=16384 7-mers. On average the chances that any given 7-mer would occur once in an rRNA sequence is one in ten. In a set of 35 rRNA sequences, a typical 7-mer would normally occur once in three or four

Biomedical Sciences Today
An open access peer reviewed journal
MDT Canada press
http://www.mdtcanada.ca/bmst.html

randomly selected sequences and none in the rest. Similarly, the chances that any 8-mer would occur is one in forty. In comparison, on average each 4-mer would occur about six times in an rRNA sequence and each 5-mer one and one-half times. That n-trees are phylogenetically fair when n≥7 but poor when n≤5 infers two things: First, n-mers that occur with frequencies close to the average frequency are not useful for taxonomy, whereas n-mers that are highly overrepresented are. Second, overrepresented of n-mers in rRNA sequences are not chance happenings; the number of overrepresented n-mers common to rRNA sequences is so large that they constitute an important expression of taxonomy, as is manifest in the existence of more than six hundred ER7s. Incidentally, this phenomenon helps to explain why the method of oligonucleotide catalog [14, 35-37] was taxonomically useful. In this method a partial list of n-mers occurring in a sequence is obtained through multiple cleavage of the sequence by nucleases, and distances between pairs of sequences are computed from the partial list [35,36].

4.2 Evolution-related n-mers dominate the distant matrix

Table 4 clearly shows overrepresented 7-mers, and by inference, other overrepresented n-mers, play a dominant role in expressing major bifurcations in the evolutionary tree. It is highly unlikely that the pattern of frequency occurrence seen in any one of the columns in the table occurred by chance. Although only a small set of the ER7s are listed in Table 4, they already bring forth the clear division of the three domains. The table also show more subtle branchings. For example, the Bacteria-Archaea ambivalence of the two thermotogales *A. aeolicus* and *T. maritima* is manifest in the frequency patterns of the ER7s GCACAAG (set 9), AATTCGA (set 13), CAGGCGC (set 16), and CTTGTAC (set 17). GCACAAG and CTTGTAC are Bacteria ER7s (category IV) but are absent in *A. aeolicus*; AATTCGA is a Bacteria ER7 but is absent in *T. maritima*; CAGGCGC is an Archaea and Eukarya ER7 (category I) but also occurs in *A. aeolicus* while absent in all other bacteria. Table 4 also gives mixed signals. For example, the closely related pair *M. genitalium* and *M. pneumoniae* are alone among the bacteria in having the Archaea and Eukarya ER7 AAACTTA (set 10) and the three Eukarya ER7s (category II) TTTGACG, TTAAAAA, and AGGGTTC (set 19). This may be a hint of a more complex genealogical relation between the two organisms and Archaea than can be suggest by any tree. In spite of these two seeming misplacements the two organisms are firmly place within Bacteria in all the best n-trees.

We can now answer the first part of the third question posed in the introduction: There are preferred n-mers that play a dominant role in molecular based phylogeny. In the case of phylogeny based on the rRNAs, they are the EROs.

4.3 Some conserved words are fully conserved across domains

The conservation properties of 16S rRNA sequences has been investigated by many authors [38], however, no fully matched words that are conserved in species spanning a group as large as a domain has been identified until now.

In *E. coli* the 13-base CW GGATTAGATACCC is located on an end loop near the α-helix H24 [39], an active center responsible for subunit association of the ribosome molecule [40]. The helix is a P site tRNA footprint and H24(791) and H24(793) are IF-3 (initiation factor) footprints [41]. The fact that the word is fully conserved in species spanning the two domains Archaea and Bacteria suggests that its earliest existence should have predated the first major branching of the universal phylogenetic tree and that subunit association may be one of the first important events in the evolution of the primitive translation apparatus.

The 9-base CW AAACGAGCG, located on the helix H35 in *E. coli* is conservative in Archaea and Bacteria. Interestingly, an expansion of its 8-base subsequence AACGAGCG, the 32-base TGTTGGGTTAAGTCCCGCAACGAGCGCAACCC, is conservative in Bacteria but not in Archaea. This suggests that the expansion occurred after the diversion of Bacteria. The 9-base CW AAACTCAAA, also conservative only in Bacteria, is located between two helices, H27 and H2, where H2(912) and H2(912-915) are mutation sites causing resistance to streptomysin and footprint sites for streptomysin, respectively [42]. The archaeal phylogenetic tree at its root is divided into two major lineage, Crenarchaeota being one of them [43]. Table 5 gives three CWs that are conserved in this lineage. Their significance is not known.

The structural information mentioned above refers to the 16S rRNA in *E. coli*. As no similar information on archaeal 16S rRNA is available at

present, nothing much concrete may be said about the significance of the CWs in Archaea. In any case the structural information on the CWs given here is very incomplete, and nothing has been said about the functional significance of the CWs.

In conclusion, it has been shown that frequencies of oligonucleotides seven to nine bases long carry sufficient information that is useful for phylogeny studies and that the frequencies of a relatively small set oligonucleotides - the EROs - are highly correlated with the distance matrix used to construct the phylogenetic tree. Using the EROs we identified oligonucleotides - CWs - that are fully conserved in the rRNAs of a large set of organisms that in some cases seem to span an entire domain or two domains. One may assume that these most widely conserved CWs are likely also among the most ancient oligonucleotides in the rRNA sequences, and conjecture that these CWs play basic roles in the structure and function of early rRNA, or its precursor the primitive translational apparatus. These notions remain to be pursued further.

Conflict of Interests
The authors declare no conflict of interests regarding the publication of this article.

Acknowledment
This work was partly supported by grants from the National Science Foundation to LFL and the National Science Council (NSC 93-2311-B-008-006) to HCL.

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

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

[3] Burge, C., Campbell, A.M. and Karlin, S., Over and underrepresentation of short oligonucleotide in DNA sequences, 1992, Proc. Natl. Acad. Sci. USA 89, 1358-1362.

[4] Luo, L.F. and Ji, F.M., The preferential mode analysis of DNA sequence, 1997, Journal of Theoretical Biology 188, 343-353.

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

[6] Hao B.L., Zhang S.Y. and Lee, H.C., Fractal related to long DNA sequences and complete genomes, 2000, Chaos, Solitons and Fractals 11, 825-836.

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

[8] Smith, H. O., Tomb, J.-F., Dougherty, B. A., Fleischmann, R. D. and Venter, J. C., Frequency and distribution of DNA uptake signal sequences in the *Haemophilus influenza* Rd genome, 1995, Science 269, 538-540.

[9] Karlin, S., Mrazek, J. and Campbell, A.M., 1996, Nucl. Acid Res. 24, 4263-4272.

[10] Bakkali, M., Chen, T.Y., Lee, H.C. and Redfield, R.J., Evolutionary stability of uptake signal sequence in the Pasteurellaceae, 2004, Proc. Natl. Acad. Sci. USA 101, 4513-4518.

[11] Trifonov, E.N. and Brendel., Gnomic - A dictionary of genetic code. (Balaban, Philadelphia, 1986).

[12] van Helden, J., Andre, B. and Collado-Vides, J., Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies, 1998, J. Mol. Biol. 281, 827-842.

[13] Bussemaker, H.J., Li, H. and Siggia, E.D., Building A Dictionary for Genomes: Identification of Presumptive Regulatory Sites by Statistical Analysis, 2000, Proc. Natl. Acad. Sci. USA 97, 10096-10100.

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

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

[16] GenBank, http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/micr.html.

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

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

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

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

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

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

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

[24] Altschul, S.F., et al., Basic local alignment research tool, 1990, J. Mol. Biol.,215, 403-410.

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

[26] Moores, A.O. and Redfield, R.J., Digging up the roots of life, 1996, Nature 379, 587-588.

[27] Doolittle, R.F. et al., Determining divergence times of the major domains of living organisms with a protein clock, 1996, Science 271, 470-477.

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

[29] Doolittle, F., Fun with genealogy, 1997, Proc. Natl. Acad. Sci. USA 94, 12751-12753.

[30] Luo, L.F., et al., Search for Evolution-Related-Oligonucleotides and Conservative Words in rRNA Sequences, 2003,IEEE Proc. Comp. Sys. Bioinformatics 2003, 468-469.

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

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

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

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

[35] Fox, G.E., Peckman, K.J. and Woese, C.R., Comparative cataloging of 16S rRNA: molecular approach to prokaryotic systematics, 1977, Int. J. Syst. Bacteriol. 27, 44-57

[36] Fox, G.E., et el., Classification of methanogenic bacteria by 16S ribosomal RNA characterization, 1977, Proc. Natl. Acad. Sci. USA 74, 4537-4541.

[37] Woese, C.R. and Fox, G.E., Phylogenetic structure of the prokaryotic domain: The primary domains, 1977, Proc. Natl. Acad. Sci. USA 74, 5088-5090.

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

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

[40] Lewin, B., *Gene 5* Ch 9 (Oxford Univ. Press, 1995).

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

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

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