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

Regular Article

Oligo-distance: A Sequence Distance Determined by Word Frequencies

L. C. Hsieh[1], C.-Y. Tseng[2], Liaofu Luo[3], Fengmin Ji[4], HC Lee*[2,5]
[1]Institute of Information Science and Genomics Research Center Academia Sinica, Taipei 115, Taiwan
[2]Department of Physics, National Central University, Chungli 320, Taiwan
[3]Department of Physics, Inner Mongolia University, Hohhot, 010021, China
[4]School of Life Science, Northern Jiaotong University, Beijing, China
[5]Department of Life Sciences, National Central University, Chungli 320, Taiwan
*Correspondence, Email: hclee12345@gmail.com

Received 2 March 2015; Accepted 19 March 2015; Published 10 April 2015

Abstract
Differences in the frequencies of chemical words of a given length in two nucleic sequences are used to define an "oligo-distance" between the sequences. Oligo-distances are much easier and faster to compute than the distances conventionally determined by sequence alignment. A correlation between oligo-distance and alignment-distance is observed. The two kinds of distances are used to construct phylogenetic trees for artificially generated sequences and for a set of thirty-five 16S and 18S rRNA sequences. The gross topologies of the trees given by the two kinds of distances are identical when the sequences are complete but only the oligo-distance is robust against sequence deformations such as rearrangement, truncation and random concatenation.

Keywords: DNA, rRNA, sequence distance, oligonucleotide occurrence frequency, sequence alignment, phylogeny, evolution.

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

1. Introduction

Sequence alignment is a standard method for comparing biological sequences. This method has proved very useful in computational molecular biology and molecular evolution. Woese [1] employed it on the 16S/18S rRNA sequences of a large and inclusive set of organisms, and discovered the three-domain - Bacteria, Archaea and Eukarya - classification. Notwithstanding impressive success of his approach, phylogenetic trees constructed based on different gene sequences may not be compatible [2] . Indeed, "no consistent organismal phylogeny has emerged from individual protein phylogenies so far produced [3]."

Alignment based distances are highly dependent on details of the alignment algorithm used and the sequence set chosen. As well, an entirely objective multiple alignment of a set of distantly related long sequences is not feasible. These limitations call for an alternative definition of evolutionary distances that are computationally practical for comparing long sequences.

Here we define an "oligo-distance" based on the difference between frequency distributions of n-mers - oligonucleotides n chemical letters long - in two sequences and use it to construct phylogenetic trees. The length n is to be judiciously chosen to produce the best effectiveness of the oligo-distance. An oligo-distance between two sequences is uniquely determined by n, and the computation is easily carried out. It has been pointed out that short range correlations among nucleotides in a DNA sequence carry evolutionary information [4-7]. We will show there is a close correlation between oligo-distance and alignment based distance.

Before applying the oligo-distance method to very long sequences, here we report the results of several sets of tests on the method using relatively short sequences, by comparing oligo-distances with alignment bases distances. First, we compare the two kinds of sequences on tree construction on a set of sequences whose phylogenetic relation we already know. The sequences, about 1600 letters long, are artificially computer generated by random bifurcation and mutation. Both methods succeed in reconstructing the original tree. Second, we repeat the test but now using a set of thirty-five 16S rRNA (for archaeons and bacteria) and 18S rRNA (for eukaryotes) sequences. The trees obtained from the two methods are of similar quality. Third, we separately rearrange the 16S rRNA sequences, truncate them, and concatenate them in random order with 23S and 5S rRNA sequences and repeat the second test. The oligo-distance trees are robust against such sequence transformations but the alignment-distance trees are not.

2. Methods

2.1 Artificial sequences for the control tree

We generate a set of 32 artificial sequences and an artificial control tree by causing a randomly chosen initial sequence 1600 bases long to bifurcate 60 times. At each bifurcation random point mutations are enacted on 1% of the two progenies. The point mutations include replacement, insertion and deletion. After the fifth bifurcation the number of progenies is limited to 32 by random pruning.

Table 1. The 35 organisms, their single-letter or symbol codes and the accession numbers of the DNA sequences of their 16S/5S/23S rRNA genes in the Genbank. For eukaryotes 18S instead of 16S rRNA. When only one accession number is given it is for the 16S rRNA.

Code	Organism	16S/5S/23S rRNA accession number(s)
A	*Aeropyrum pernix*	AB019522/AP000062/AB019552
B	*Pyrococcus horikoshii*	D45214
C	*Archaeoglobus fulgidus*	Y00275/AE000782/AE000782
D	*Methanococcus jannaschii*	M59126/L77117/L77117
E	*Methanobacterium thermoautotrophicum*	Z37156
F	*Thermoproteus tenax*	M35966
G	*Methanothermus fervidus*	M32222
H	*Sulfolobus solfataricus*	X03235/X01588/U32322
L	*Halobacterium volcanii*	D11107
a	*Escherichia coli*	Z83204/AE000461/AE000461
b	*Haemophilus influenzae*	M35019/U32847/U32847
d	*Helicobacter pylori*	U00679/U27270/U27270
e	*Rickettsia prowazekii*	M21789
f	*Bacillus subtilis*	AF058766/Z99119/Z99119
g	*Mycoplasma genitalium*	X77334/U39694/U39694
h	*Mycoplasma pneumoniae*	M29061/AE000007/X68422
i	*Mycobacterium tuberculosis*	X52917/Z73902/Z73902
j	*Synechococcus sp.*	D90916/D90916/D90916
k	*Borrelia burgdorferi*	X98233/X57791/M88330
m	*Treponema pallidum*	M88726/AE000520/AE000520
n	*Chlamydia trachomatis*	D85720/AE001347/AE001347
o	*Chlamydia pneumoniae*	L06108/AE001363/AE001363
p	*Flavobacterium heparinum*	M11657
q	*Deinococcus radiopugans*	Y11334
r	*Herpetosiphon aurantiacus*	M34117
s	*Chlorobium limicola*	Y08102
y	*Aquifex aeolicus*	AE000657/AE000657/AE000657
z	*Thermotoga maritima*	AE001703/AE001703/AE001703
%	*Homo sapiens* (human)	M10098
!	*Mus musculus* (mouse)	X00686
@	*Solanum tuberosum* (potato)	X67238
*	*Glycine max* (soybean)	X02623
#	*Drosophila melanogaster* (fly)	M21017
$	*Caenorhabditis elegans* (worm)	X03680
&	*Saccharomyces cerevisiae* (yeast)	J01353

2.2 rRNA sequences

The 16S/18S rRNA sequences of 35 organisms - 9 archaeons, 19 bacteria and 7 eukaryotes - are downloaded from the GenBank [8]. The names of the organisms and the accession numbers of the rRNA sequences are listed in Table 1. The organisms are selected according the twin criteria of coverage and availability. In Table 1 each archaeon is coded by an upper-case Roman alphabet, each bacterium by a lower-case alphabet and each eukaryote by a non-alphabet symbol.

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

2.3 The oligo-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 $\sigma = abc\cdots$ is an n-mer, denote the joint probability, or relative frequency of the *n*-mer σ occurring in the sequence, by p_σ. We obtain p_σ as follows [9]. We use a sliding window of width n to count the frequencies of occurrence f_σ for all the 4 n (overlapping) *n*-mers σ in the set {σ} n and get $p_\sigma = f_\sigma / N$, where *N* is the sequence length. Operationally we treat each sequences as if were circular so that for any n the sum-rules $\sum_\sigma f_\sigma = N$ and $\sum_\sigma p_\sigma = 1$ hold. When 4^n is less than the sequence length *N*, the set $\{p_\sigma\}$ with increasing *n* is an increasingly fine-grained characterization of a sequence. When 4^n is greater than *N*, at most *N* of 4^n *n*-mers have nonzero probability and, with increasing n, the set $\{p_\sigma\}$ will increasingly be an expression of overrepresented *n*-mers.

Given two sequences α and β with joint probability sets $\{p_{\alpha,\sigma}\}$ and $\{p_{\beta,\sigma}\}$, respectively, the *n*-distance between the two sequences is defined as [10]

$$D_{\alpha\beta}^{(n)} = \sum_{\sigma\in\{\sigma\}_n} \left| p_{\alpha,\sigma} - p_{\beta,\sigma} \right|, n = 1,2,... \qquad (1)$$

An *n*-distance is well defined for sequences that are of different lengths and are not aligned. For the moment write $D(\alpha\beta) \equiv D_{\alpha\beta}^{(n)}$ for any *n*, then $0 \le D(\alpha, \beta) \le 2$ for all sequences α and β. Furthermore, *D* satisfies the formal requirements for a distance: (a) $D(\alpha, \alpha)=0$, (b) $D(\alpha, \beta)>0$, (c) $D(\alpha, \beta)=D(\beta, \alpha)$, (d) $D(\alpha, \beta)+D(\beta, \alpha)> D(\alpha, \gamma)$ for any α, β and γ. The computation time required for an *n*-distances grows linearly with sequence length whereas the computation time required for a distance based on sequence alignment grows exponentially with sequence length. For a set of *Q* sequences, we define a *Q*×*Q* *n*-distance matrix $D^{(n)}$ whose elements are the pairwise *n*-distances defined in Eq. (1), and an *n*-similarity matrix $S^{(n)}$ defined as

$$S^{(n)} \equiv 1 - \frac{1}{2} D^{(n)} \qquad (2)$$

Here the unity symbol stands for the unit matrix. The n-similarity between two sequences α and β, that is the element $S_{\alpha\beta}^{(n)}$, is equal to 0 when α and β are totally dissimilar, and to 1 when they are identical. Both $D^{(n)}$ and $S^{(n)}$ are symmetric matrices.

Suppose *Q* is the number of sequences in a set of sequences, *N* is the mean length of the sequences, and the variance in the lengths of the sequences is not large. Then in order to give useful oligo-distances n must be such that $4^n >> Q$. There is not an obvious constraining relation between *n* and *N*. When $4^n << N$ each *n*-mer will on average occur many times in the sequence and the *n*-distance will be weighed more by *n*-mers occurring with close to the average frequency then by overrepresented *n*-mers. In this case the n-distances will have too similar values to be useful. As *n* increases the weighing of the *n*-distance will be shifted towards overrepresented *n*-mers and at some point an n-distance of maximum utility is expected to obtain. When 4^n approaches infinity (relative to *N*) all sequence will be completely distinct and the *n*-distance becomes useless. We therefore expect the optimal n to be such that $4^n \ge N$. The lengths of the 16S/18S rRNA sequences are about 1500 to 1800 letters and lie in the range 4^5 =1024 and 4^6 =4096. So a first guess is that the optimal *n* would be about 6. As it turns out the optimum *n* value is 8 or 9.

2.4 Tree construction

The procedure we use for tree construction depends on whether we use the oligo-distance method or use sequence alignment. In the case of oligo-distance, we decide on an *n* value and use the procedures described above to compute the $D^{(n)}$ and $S^{(n)}$ matrices. In the case of sequence alignment, we use the software package OMIGA 1.13 [11] to make multiple alignment of the set of sequences and to generate an alignment similarity matrix (X matrix). In both cases we use the neighbor-joining (NJ) method (see, for instance, Li [12]) and the package CLUSTL16 [13] to construct a tree, or a dendogram; an oligo-tree from an S matrix, and an alignment-tree from an *X* matrix.

3. Results

3.1 Results from artificial sequences

We generated many sets of artificial sequences and from each set we used the oligodistance method to construct $S^{(n)}$ matrices for *n*=2 to 9 and the alignment method to construct an *X* matrix. A strong and persistent correlation between corresponding elements of the *S*, for *n*≥7, and *X* matrices was observed. Fig.1 shows a typical case for *n*=9. In the figure, each piece of data gives the logarithms of $S^{(9)}$ (y-axis) and *X* (x-axis)

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

similarities between a pair of artificial sequences. The observed correlation can be expressed by the relation

$$S_{\alpha\beta}^{(n)} \approx \left(X_{\alpha\beta}\right)^{5.5\pm0.5} \quad (3)$$

holds for elements of the $S^{(n)}$ and X matrices (a discussion of the origin of this relation will be presented elsewhere [14]). We thus define a hatted S matrix $\hat{S}^{(n)}$ by

$$\hat{S}_{\alpha\beta}^{(n)} \approx \left(S_{\alpha\beta}^{(n)}\right)^{1/5.5} \quad (4)$$

and a hatted distance matrix $\hat{D}^{(n)} \equiv 2\left(1-\hat{S}^{(n)}\right)$. Then the hatted S and D matrix elements will be approximately linearly related to their counterparts in the alignment method. We observe that the approximate linear relation between log S and log X begins to show distortions when $X<0.45$ or when $S<0.02$. That is, when the similarity between two sequences is weak. For reference, the alignment-similarity between two random sequences 1600 bases long is about 0.4.

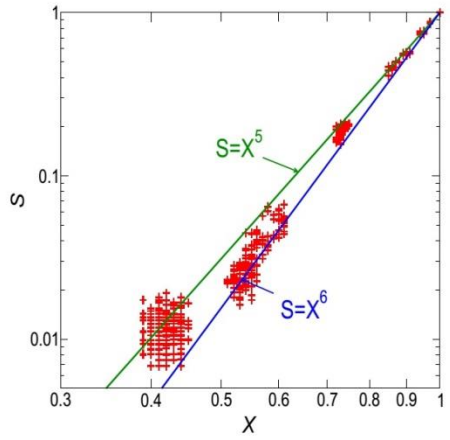
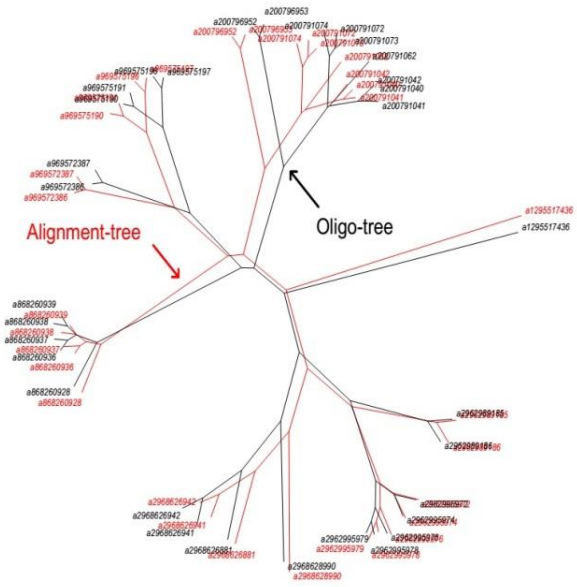


Figure 1. Log-Log plot of oligo-similarity (S) versus alignment-similarity (X). Each point in the figure is the similarity for a pair of artificial sequences on a 32-node control tree. See text for the generation ofartificial sequences. Straight line gives the relations $S = X^5$ and $S = X^6$.

Figure 2 shows two phylogenetic trees constructed respectively from the $\hat{S}^{(9)}$ (black) and X (red or dark gray) matrices. The trees are topologically identical and phylogenetically correct. The bunching of data in Fig. 1 can now be understood. The sequences are grouped in clades, and the bunching of data in Fig. 1 is a reflection of this grouping. That is, bunching forms when the variance in intraclade similarity is smaller than the difference in the average interclade similarity. The result is similar when n=7 and 8 but the quality of the oligo-tree deteriorates with decreasing n when $n\leq6$. To summarize, this part of the study suggests that for the simple evolution events used to generate control trees, the oligo-distances for some n are as good as alignment based distances.

3.2 Oligo-tress from rRNA sequences

We used the oligo-distances to compute pairwise similarities $S^{(n)}$, n= 2 to 9, for the 16S/18S rRNA sequences of the 35 organisms listed in Table 1. These were converted to $\hat{S}^{(n)}$ through Eq. (4). The hatted $S^{(n)}\left(D^{(n)}\right)$ matrices were then used for constructing oligo-trees for the 35 organisms. Fig. 3 shows the unrooted trees constructed from the hatted n-distances, n = 2, 3, 4, 5, 7, 8. In the figure, each archaeon is coded by an upper-case Roman alphabet, each bacterium by a lowercase alphabet and each eukaryote by a nonalphabet symbol. The qualities of the trees are poor for $n \leq 4$. When n = 5 and 6 (not shown) the Archaea and Eukarya domains are roughly separated but both appear as branches within Bacteria. The three domains - Bacteria, Archaea, and Eukarya - are separated almost correctly when n = 7 except that two hyperthermophile, Aquifex aeolicus (y) and Thermotoga maritima (z), are grouped with Archaea instead of Bacteria. When n=8 and 9 (not shown), the three domains are completely separated, and A. aeolicus and T. maritima are correctly shown to be the two deepest branching bacteria [15-17].



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

Figure 2. Reconstructed trees from artificial sequences using n = 9 oligo-distance (black) and distance based on sequence alignment (red). The two trees are topologically identical and correctly reproduce the topology of the original 32-node artificially generated control tree.

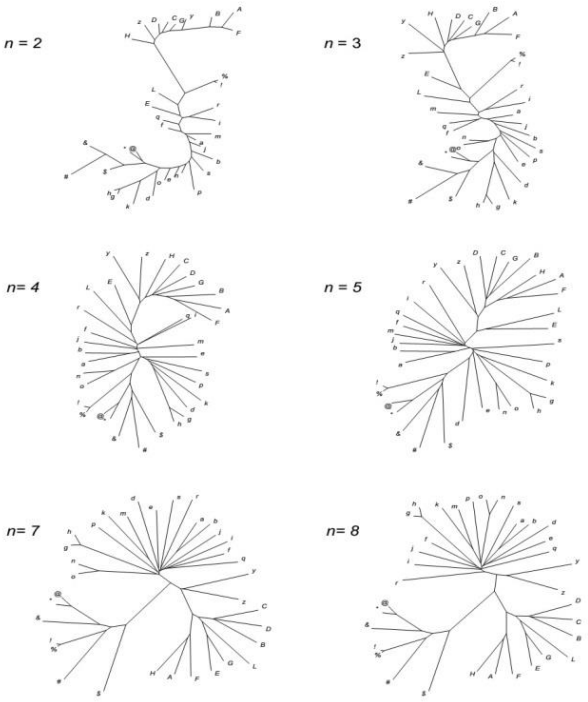


Figure 3. Unrooted oligo-trees constructed for the 35 organisms using different n-distances. The three domains - Bacteria, Archaea, and Eukarya - are completely separated when n = 8, and *A. aeolicus* (y) and *T. maritima* (z) are correctly shown as the two deepest branching bacteria.

3.3 Correlation between S and X for rRNA sequences

The alignment based similarity matrix X for the 35 rRNA sequences are computed using CLUSTL16. In Fig. 4 the logarithms of elements of the $S^{(9)}$ are plotted against those of X. The data shows the relation $S^{(9)} \approx X^{5.5\pm0.5}$ holds well for X>0.54 or S>0.5. This is a smaller range of validity for the relation than seen in Fig. 1. One possible cause for this discrepancy is the difference in the way the artificial sequences and the rRNA sequences diverged. Whereas point mutations were the only way the artificial sequences were made to diverge, many other modes of mutation, including relatively large deletions and insertions and translocations, must have also contributed to the divergence of rRNA sequences. Aside from a small number of points near $X \approx S \approx 1$, data points in Fig. 4 appear in three large clusters. Data in the $0.65 \lesssim X \lesssim 0.85$ cluster mostly come from intradomain pairs, in the $0.54 \lesssim X \lesssim 0.65$ cluster mostly from Bacteria-Archaea pairs, and in the $X \lesssim 0.54$ mostly from Bacteria-Eukarya and Archaea-Eukarya pairs.

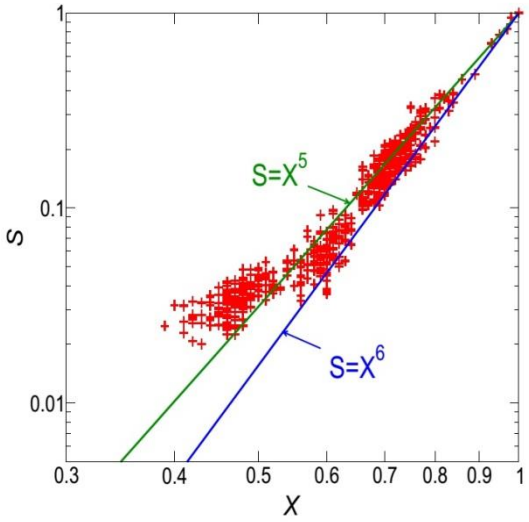


Figure 4. Log-Log plot of oligo-similarity ($S^{(9)}$) versus alignment-similarity (X). Each point in the figure gives the similarities between the rRNA sequences from a pair among the 35 organisms in Table 1. Straight lines give the relations $S^{(9)} = X^5$ and $S^{(9)} = X^6$.

3.4 Tree of Life

Fig. 5 shows the 35-organism tree obtained using respectively the $\hat{S}^{(9)}$ (black) and X (red or dark gray) matrices. The two trees have gross topologies that are similar to the Tree of Life 18 with all branches not leading to one of the 35 organisms pruned. On both trees, the three domains are cleanly separated, the two hyperthermophiles *A. aeolicus* (y) and *T. maritima* (z) are correctly shown as the two deepest branching bacteria, the archaeons are correctly separated into Euryarchaeota and Crenarchaeota, and *C. elegant* (worm, $) is incorrectly given as the deepest branching eukaryote instead of the two plants *G. max* (soybean, ∗) and *S. tuberous* (potato, @) 18,19. All the closely related organisms are correctly paired on both trees: *E. coli* (a) and *H. influenza*(b), *M. genitalia* (g) and *M. pneumonia* (h), and *Ch. trachomatis* (n) and *Ch. pneumonia* (o) among the bacteria, and *H. sapiens* (%) and *M. musculus* (!), and *S. tuberous* and *G. max* among the eukaryotes.

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

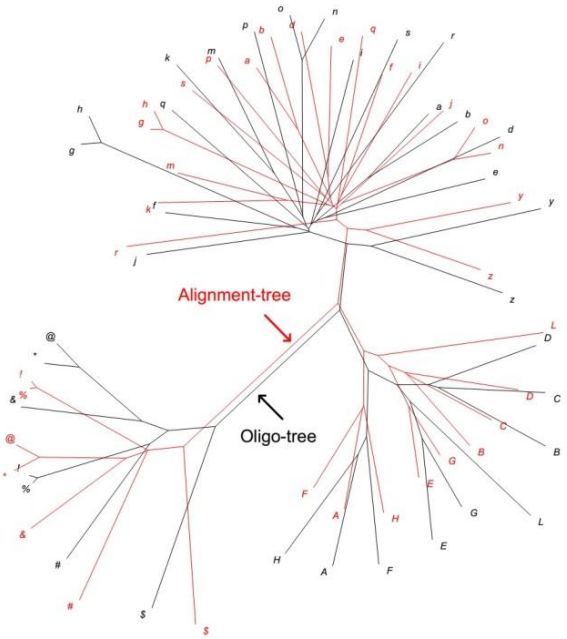


Figure 5. Tree of Life for the 35 organisms listed in Table 1. Red or dark gray tree is constructed from alignment-distances and black tree is from *n* = 9 oligo-distances.

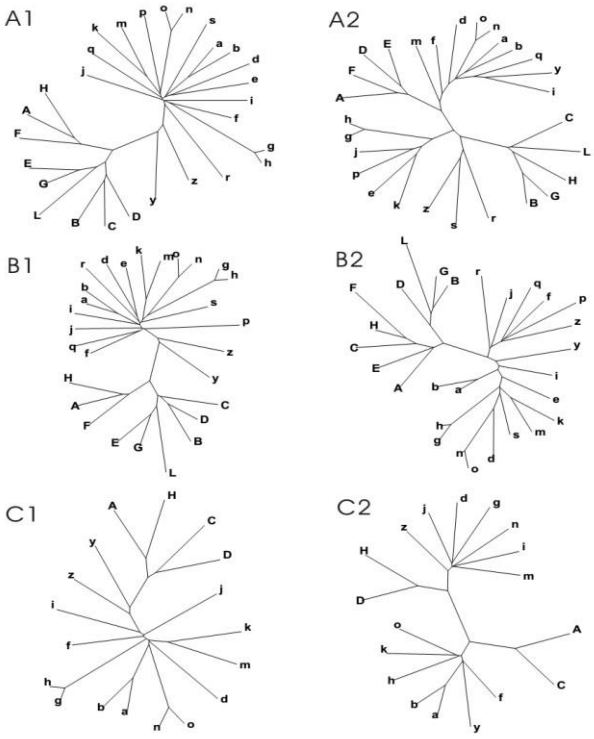


Figure 6. Oligo-trees (left) and alignment-trees (right) constructed using transformed or new rRNA sequences. (A) A 16S rRNA sequence is spliced at a random site and the two resultant segments are transposed and reconnected. (B) A 16S rRNA sequence is randomly truncated to a segment 800 to 1200 bases long. (C) For each organism a new sequence is formed by concatenating in random order the 5S, 16S and 23S rRNA sequences.

3.5 Robustness of oligo-tree

We tested the robustness of oligo- and alignment-trees against three types of transformations and alterations done to the rRNA sequences. Only the 19 bacteria and 9 archaeons are used in the tests. In each case only the oligo-tree is robust against the transformation. In the first test each 16S rRNA sequence is spliced at a randomly chosen site and a new sequence is made by transposing and reconnecting the two resultant segments. The transposed sequences are then used to constructed an oligo-tree (Fig. 6, A1) and an alignment-tree (Fig. 6, A2) with methods employed to obtain the trees in Fig. 5 (but without the eukaryotes). It is seen that the oligotree has the correct topology - Bacteria and Archaea are cleanly separated and the two hyperthermophiles *A. aeolicus* (y) and *T. maritima* (z) are the two deepest branching bacteria - while the alignment-tree does not - the archaeons are incorrectly broken into two groups that are admixed with bacterial groups. This test shows that the oligo-distance is robust against sequence rearrangement. In the second test each 16S rRNA sequence is randomly truncated - possibly from both ends - to a segment 800 to 1200 bases long. The resulting oligo-tree (Fig. 6, B1) again has the correct topology whereas on the alignment-tree (Fig. 6, B2) the Archaea and Bacteria are separate but both are incorrectly partitioned. This test suggests that the oligo-distance may be useful for phylogeny even when only fragments of sequences are available. In the third test, for each organism a sequence (about 3500 bases long) is formed by concatenating in random order the 16S, 23S (2840 to 3030 bases long), and 5S (105 to 130 bases long) rRNA sequences. In this test the set of organisms is reduced to 14 bacteria and 4 archaeons (Table 1) by the availability of the three rRNA sequences. Again the topology of the oligo-tree is correct but not so the alignment-tree. On the latter the archaeons are not separated from the bacterials, the two pairs of archaeons each has one euryarchaeote and one crenarchaeote, and the two hyperthermophiles *A. aeolicus* (y) and *T. maritima* (z) are seriously misplaced. This test again shows that the oligo-distance is insensitive to sequence rearrangement. It furthermore shows the 23S rRNA sequences, and to a lesser extent the 5S rRNA sequences, yield oligo-distances that are consistent with those given by 16S rRNA sequences.

4. Discussions

Our experiments with artificial and real DNA sequences indicate that the oligodistance is a promising tool for phylogeny. We found that all oligo-distances are not the same and, in the studies conducted here with rRNA sequences averaging about 1600 bases long, *n*-distances with *n*=8 and 9 are the best. For a sequence 1600 bases long, the nature of an 8-distance is fundamentally different from that of, say, a 5-distance. Because 1600 is of the order of magnitude of 45=1024, a 5- distance is mainly determined by the majority of 5-mers, which necessarily have close to the average frequency of occurrence. In contrast, 48=65,536 is about forty times 1600, the vast majority of 8-mers have zero frequency of occurrence and the 8-distance is determined by highly overrepresented 8-mers, that is, those 8-mers whose frequencies of occurrence are equal or greater than forty times the average frequency. From the fact that we get good quality oligo-trees only when *n*=8 and 9 we infer that phylogenetic relations between DNA sequences are reflected through the small subset of *n*-mers that are highly overrepresented. Further consequences of this inference are being investigated.

Oligo-distance is devised to supplement the sequence alignment method, not to replace it. Sequence alignment is still by far the best method to detect small differences in sequences that have a high degree of similarity. A surprise in our investigation is the power-law relation between the oligo- and alignment-similarities. This relation guarantees that trees generated using oligo-distances and alignmentbased distances resemble each other. However, for it to hold true, the correlation has a higher required minimum similarity for the rRNA sequences (X=0.54, Fig. 4) than for the artificial sequences (X=0.45, Fig.1). This difference causes the oligoand alignments-trees to be identical for the artificial sequences but to differ in the details for the rRNA sequences. The exponent of 5.5 in Eq. (3) can be partially derived for homologous sequences that diverged by point mutations 14. An early form of this correlation was implicit in Fig. 2 in Woese's 1987 paper 1, between what Woese called the binary association coefficient and alignment-similarity for 16S rRNA sequences. The binary associate coefficients were computed from *n*-mer catalogs determined in wet laboratories [20], and involved *n*-mers of various lengths, all of which were longer than eight letters and many were much longer.

Although oligo-distance succeeds in generating a (partial) Tree of Life that is correct in its gross topology, it fails in getting the finer details correctly. Thus modifications to the simple procedure described here needs to be made before the method can be considered for detailed phylogeny studies. In this respect we are encouraged by the recent report of a successful application of oligo-distance for *peptide* sequences to the whole-genome phylogeny of prokaryotes [21]. There are also other non-alignment based methods that have been devised for sequence comparison and applied to (small scale) phylogenetic tree construction using long sequences, including mitochondria complete genomes [22,23].

We showed oligo-distances has the distinctive feature that trees constructed from it are robust against rearrangement, truncation, and concatenation of sequences. This makes *n*-distances potentially useful for phylogeny in a variety of situations, including the two opposite ones: when only different collections of relatively short fragments are available for the different species to be compared, and when large collections of genes or even complete genomes are available.

Acknowledgement
This work is partly supported by grants 92-2112-M-008-040 and 93-2311-B-008-006 from the National Science Council (ROC) to HCL.

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

References

[1] Woese, C.R. Bacterial evolution. *Microbial. Rev*., **1987**, 51, 221–271.
[2] Baldauf, S.L.; Palmer, J.D. and Doolittle, W.F. The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. *Proc. Natl. Acad. Sci. USA*, **1996**, 93, 7749–7754.
[3] Woese, C.R. The universal ancestor. *Proc. Natl. Acad. Sci. USA*, **1997**, 95, 6854-6859.
[4] Burge, C.; Campbell, A.M. and Karlin, S. Over and under-representation of short oligonu-cleotide in DNA sequences. *Proc. Natl. Acad. Sci. USA*, **1992**, 89, 1358–1362.
[5] Lou, L.F.; Ji, F.M. and Li, H. Fuzzy classification of nucleotide sequences and bacterial evolution. *Bull. Math. Biol*., **1995**, 57, 527–537.
[6] Karlin, S.; Mark, J. and Campbell, A.M. Compositional biases of bacterial genomes and evolutionary implications. *J. Bacteriology*, **1997**, 179, 3899–3913.
[7] Lou, L.F.; Lee, W.; Jia, L. et al. Statistical correlation of nucleotides in a DNA sequence. *Phys. Rev. E*, **1998**, 58, 861–871.

[8] All sequences are downloaded from GenBank at http://www.ncbi.nlm.nih.gov/
[9] Hao, B.L.; Zhang, S.Y. and Lee, H.C. Fractal related to long DNA sequences and complete genomes. *Chaos, Solitons and Fractals*, **2000**, 11, 825-836.
[10] Luo, L.F.; Hsieh, L.-C.; Ji, F. et al. Search for Evolution-Related-Oligonucleotides and Conservative Words in rRNA Sequences. *IEEE Proc. Comp. Sys. Bioinformatics CSB'03*, **2003**, 468-469.
[11] Calvet, J.P. Comprehensive sequence analysis: OMIGA 1.1. *Science*, **1998**, 282, 1057–1058.
[12] Li, W.H. Molecular Evolution., Signer Associates, **1997**.
[13] Thompson, J.D.; High, D.G. and Gibson, T.J. Improved sensitivity of profile searches through the use of sequence weights and gap excision. *Computer Appl. in Biosciences*, **1994**, 10, 19–29.
[14] Lee H.C. et al. (unpublished).
[15] Achenbach-Richter, L.; Gupta, R.; Setter, K.O. and Woese, C.R. Were the original eubacteria thermophiles? *Syst. Appl. Microbial*., **1987**, 9, 34-39.
[16] Burggraf, S.; Olsen, G.J.; Stetter, K.O. and Woese, C.R. A phylogenetic analysis of Aquifex pyrophilus. *Syst. Appl. Microbial*., **1992**, 15, 352-356.
[17] Nelson, K.E.; Calyton, R. A.; Gill, S. R. et al. Evidence for horizontal gene transfer between archaea and bacteria from genome sequence of T. maritima. *Science*, **1999**, 399, 323-329.
[18] Patterson, D.J. and 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., 1993) pp. 13-46. See also the "Tree of Life" website: [phylogeny.arizona.edu/tree/eukaryotes/crown eukaryotes.html].
[19] Feng, D.F.; Cho, G. and Doolittle, R.F. Determining divergence times with a protein clock: Update and reevaluation. *Proc. Natl. Acad. Sci. USA*, **1997**, 94, 13028-13033.
[20] Fox, G.E.; Peckman, K.J. and Woese, C.R. Comparative cataloging 0f 16S ribosomal ribonucleic acid: molecular approach to prokaryotic systematics. *Int. J. Syst. Bactreiol*., **1977**, 27, 44-57.
[21] Qi, J.; Wang, B. and Hao, B.L. Whole genome prokaryote phylogeny without sequence alignment: a K-string composition approach. *J. Mol. Evol*., **2004**, 58, 1-11.
[22] Chen, X.; Kwong S. and Li, M. A Compression Algorithm for DNA Sequences and Its Applications in Genome Comparison. Proc. 4th Annual Intl Conf. Comp. Molecular Biology (RECOMB4), R. Shamir et al., (Eds.) 107-117, **2000**.
[23] Li, M.; Badger, J. H.; Chen, X.; Kwong, S.l Kearney, P. and Zhnag, H. An information-based sequence distance and its application to whole mitochondrial genome. *Bioinformatics*, **2001**, 17, 149-154.