Search for Evolution-Related-Oligonucleotides and Conservative Words in rRNA Sequences

Liaofu Luo[1*], Li-Ching Hsieh[3], Fengmin Ji [2], Mengwen Jia[1], H.C. Lee[3,4*]

[1]*Physics Department, Laboratory of Theoretical Biophysics, Inner Mongolia University, Hohhot 010021, China*

[2]*Department of Physics, Northern Jiaotong University, Beijing 100044, China*

[3]*Department of Physics &* [4]*Department of Life Sciences, National Central University, Chungli 320, Taiwan*

*Corresponding authors

Abstract

We describe a method for finding ungapped conserved words in rRNA sequences that is effective, utilizes evolutionary information and does not depend on multiple sequence alignment. Evolutionary distance (called n-distance) between a pair of 16S or 18S rRNA sequences is defined in terms of the difference in the two sets of frequencies of occurrence of oligonucleotides n bases long (n-mers) given by the sequences. These n-distances are used to reconstruct phylogenetic trees for 35 representative organisms from all three kingdoms. The quality of the tree generally improves with increasing n and reaches a plateau of best fit at n=7 or 8. Hence the 7-mer or 8-mer (oligonucleotide of 7 or 8 bases) frequencies provide a basis to describe rRNA evolution. Based on the analysis of the contribution of a particular 7-mers to 7-distances, a set of 612 7-mers (called evolution-related-oligonucleotides, EROs) that are critical to the topology of the best phylogenetic tree are identified. Expanding from this set of EROs, evolution-related conservative words longer than 7 bases in 16S rRNA sequences from an enlarged set of 98 organisms in Bacteria and Archaea are identified based on two criteria: 1) *the word is highly conserved in nearly all species of a kingdom (or a sub-kingdom); and* 2) *the word is located at nearly the same site in each sequence. Three examples of words thus found are: The 13-mer* ggattagataccc *located at the end of a loop near* H24 *(in E.coli) is conservative in almost all species in Archaea and Bacteria. The 8-mer* aacgagcg *located on* H35 *is also conservative in Archaea and Bacteria. Its expansion, the 32-mer* tgttgggttaagtcccgcaacgagcgcaaccc, *is conservative in Bacteria but not in Archaea.*

1. Introduction

Evolutionary tree gives a detailed description of evolutionary relations. The requirement of a theoretically deduced tree consistent with life tree is a very strong constraint. We shall use 16S rRNA (18S rRNA) to reconstruct evolutionary tree. Instead of multi-alignment of sequences a new definition of evolutionary distance which is based on the oligo-nucleotide frequency will be proposed. In this analysis the evolutionary tree will reveal a set of conservative words in rRNA sequences and these evolution-related conserved words provide a clue to explore the evolutionary relations on translational apparatus and mechanism.

The 35 representative organisms (called set 1, including 9 archaeons, 19 bacteria and 7 eukaryotes) are studied in constructing evolutionary tree and for the purpose of further studies on conserved word a test set of 16S rRNA sequences for prokaryotes are selected which includes 61 organisms – 20 archaeons and 41 bacteria. [1].

2. Evolutionary distance and reconstruction of evolutionary tree

Let $\sigma = abc\ldots$ being an oligonucleotide n bases long. Given two sequences Σ and Σ' with sets of joint probabilities $\{p_\sigma\}$ and $\{p_\sigma'\}$, respectively, define a distance, called an n- distance, between the two sequences based on the difference of joint probabilities in the two sets

$$E_n(\Sigma,\Sigma') = \sum_\sigma \left| p_\sigma - p_\sigma' \right| \qquad (1)$$

For each n, $2 \le n \le 9$, we compute distance matrix D for the 35 organisms. Dendograms, or n-trees, are then constructed from the distance matrix using the UPGMA method, NJ method and FC (fuzzy clustering) method respectively. We have investigated how the n-tree changes with n and found that, irrelative with n-tree construction, the best n-trees are obtained at n=7 or 8. So, the oligo-nucleotide frequency gives a good definition of evolutionary distance. The success of this approach also indicates the possible existence of some preferred words with length near 7, the frequency of which correlates with evolution.

Proceedings of the Computational Systems Bioinformatics (CSB'03)
0-7695-2000-6/03 $17.00 © 2003 IEEE

3. Evolution-related oligonucleotides with n=7

For a set of 35 sequences, the distance matrix includes 35×34/2=595 nonvanishing elements. These elements are independent one another. If the summation on the right-hand-side of Eq (1) is removed and only a single term of oligonucleotide σ is retained, then a "single-word-distance" based on σ, $D_{sw}(\sigma)$, can be defined. For a sampling size of 595, the correlation coefficient Cor(σ) between n-distance D and $D_{sw}(\sigma)$ is substantially greater than the threshold value (at 99% C.L.) of 0.11 may be considered as indicating special significance for word σ in the evolution process. Among all n=7 words 612 are found to have Cor(σ)≥ 0.30. Taking the latter value arbitrarily to be the cut-off value, we call these words n=7 evolution-related- oligonucleotides (ERO7s). These ERO7s generally occur only in one or two kingdoms. So, the bifurcation of evolutionary tree is closely related to and can be described by the occurrence of these oligomers.

4. Conserved words in three kingdoms

From a set of ERO7s we find conserved words (CWs) in three kingdoms by the procedure of three steps: 1) Identify EROs equal or longer than 7 bases by matching all ERO7s with 16S (18S) rRNA sequences. Note that in match some words are partly overlapped and they should be melted each other, forming a longer word. Then we collect all obtained words in database with length equal or larger than 7. 2) By use of BLAST program check the matching sites and identify those EROs with length equal or larger than 8 as candidate CWs whose relative positions in 16S (18S) rRNA sequences are approximately fixed for a large number of organisms in set 1. 3) Identify those candidate CWs as real CWs that they also appear at the nearly same positions in a larger number of organisms in the extended set of 61+35 organisms. We have found many conservative words in above approach but the full matched words (wrong matching and the inserting /deleting are not permitted) are few. The permitted error in location in different 16S rRNAs is set to be several tens of bases. Thus, we obtain all CWs in Bacteria and Archaea. The representative kingdom-wide conservative words in Bacteria and Archaea are:

ggattagataccc (Archa / Bact.)
aacgagcg (Archa / Bact.)
gacggtgag (Archa)
ccttgcacacac (Archa)
aaactcaaa (Bact.)
tgggttaa (Bact.)
accaccag (Crenarchaeota)
gtagtcccg (Crenarchaeota)
cccgtcgc (Crenarchaeota)

The conservation of 16S rRNA sequences has been investigated by many authors. However, to our knowledge, the fully matched conserved word which is conserved in such a large range as a kingdom is firstly indicated by us. The word ggattagataccc in *E.coli* is located on end loop near H24[2]. It is an active center responsible for subunit association of the ribosome molecule. The word is highly conservative in two kingdoms — Archaea and Bacteria — of species. It transcends the era of the earliest branching of universal phylogenetic tree. So, the conservation of the word perhaps means the subunit association as the first important event in the evolution of primitive translation apparatus since the relatively rigid tRNA may be located between large and small subunits of the ribosome. Note that in *E.coli* the H24 is a P site tRNA footprint and H24(791) and H24(793) are IF-3 (initiation factor) footprint[3]. The word aacgagcg in *E.coli* is located on a helix H35. This is a 8-bases long word and also conservative in Archaea and Bacteria. Interestingly, its expansion, a 32-bases long word, tgttgggttaagtcccgcaacgagcgcaaccc, is conservative in the kingdom Bacteria. This mean probably the expansion occurring in the bifurcation of Bacteria from universal tree. Another word aaactcaaa conservative in Bacteria is located between two helices, H27 and H2, while H2(912) and H2(912-915) are mutation sites causing resistance to streptomysin and footprint sites for streptomysin [3]. All the structural information indicated above is gained by reference to the 16S rRNA of *E.coli*. Though the detailed explanation on the meaning of these conserved words has not been given one may reasonably assume that these words are closely related to the basic structure and function of bacterial ribosome, related to the early evolution of the primitive translational apparatus.

References

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

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

[3] Mueller, F., *et al.* A new model for the 3-D folding of *E.coli* 16S ribosomal RNA. J. Mol. Biol. (1997) 271, 524-544; 566-587.