International Journal of Statistical Sciences *ISSN 1683-5603*
Vol. 21(1), 2021, pp 117-126

Phylogenetic Clustering of Microbial Communities Based on 16S rRNA Sequences

Md. Mazharul Islam*, Mst. Sharmin Akter, Md. Hadiul Kabir and Md. Nurul Haque Mollah*

Laboratory of Bioinformatics, Department of Statistics, University of Rajshahi, Rajshahi 6205, Bangladesh.

*Correspondence should be addressed to Md. Mazharul Islam (Email: mazharul4787@gmail.com)

[Received September 8, 2020; Revised October 20, 2020; Accepted November 19, 2020]

Abstract

The 16S rRNA sequences are commonly using for identification, classification and quantitation of microbes in complex biological mixtures such as environmental samples like water, soil or air, and microbiome samples. These sequences are also using for phylogenetic studies as it is highly conserved between different species of microbes. Phylogenetic clustering of microbial communities based on 16S rRNA sequence is playing a vital role to identify diseases related virus and bacteria. There are several methods for the formation of a phylogenetic tree based on sequence data. Up to date, nobody compares the phylogenetic tree methods yet. In this paper, we compared the performance of several phylogenetic tree approaches including neighbor-joining (NJ), UPGMA, maximum parsimony (MP) and maximum likelihood (ML) for microbial clustering based on 16s RNA sequences collected from 11 different environments. The analyzing results of this study apprize that for microbial clustering by using a phylogenetic tree based on sequence dataset, the maximum likelihood method is comparatively better than the other three methods.

Keywords: Microbial clustering, 16S rRNA sequence, Phylogenetic tree, Maximum Likelihood approach. Diseases related to microbes.

AMS Classification: 92C42.

1. Introduction

Many biologists agree that a phylogenetic tree of relationships should be the central supporting of research in many areas of biology (Soltis, D.E. and Soltis, P.S., 2003). Comparisons of plant species of gene sequences in a phylogenetic situation can provid the most meaningful insights into biology (Hall et al., 2002a; Doyle et al., 2003). A phylogenetic tree is a diagram used to show the inferred evolutionary pathways and connections among various biological species. The 16S rRNA gene is a highly conserved component of the transcriptional machinery of all DNA-based life forms and thus is highly suited as a target gene for sequencing DNA in samples containing up to thousands of different species (Patel J. 2001; Tringe and Hugenholtz, 2008). This sequence is commonly used for identification, classification and quantitation of microbes in complex biological mixtures such as environmental samples like water, soil or air, and microbiome samples (Woese et al., 1985; Woese, C.R., 1987). The 16S rRNA gene is used for phylogenetic studies as it is highly conserved between different species of bacteria and archaea (Weisburg et al., 1991). Molecular phylogenetic analysis is the use of macromolecular sequences to reconstruct the evolutionary relationships between organisms. Phylogenetic trees represent the evolutionary relationships of sequences or species (Fitz Gibbon et al., 1999). Optimal alignment of the primary structures and a careful data selection are prerequisites for reliable phylogenetic conclusions. rRNA based phylogenetic trees can be reconstructed and the significance of their topologies evaluated by applying Neighbor-Joining, UPGMA, Maximum Parsimony and Maximum Likelihood methods of phylogeny inference in comparison, and by fortuitous or directed resampling of the data set (Horner et al., 2004). Up to date, nobody can't compare the phylogenetic tree methods yet. Only a few papers were using one of the phylogenetic tree methods among the several methods of a phylogenetic tree. In this study, we should try to explore better phylogenetic tree approach for microbial clustering based on 16s RNA sequence.

2. Materials and Methods

2.1. Data Source

We collected 16S rRNA sequence datasets from NCBI (the human gut bacteria microbiome) and European Bioinformatics Institute (EBI) Metagenomics (https://www.ebi.ac.uk/metagenomics/) from 11 different environments. Here, we consider 11 microbial environments like human gut bacteria, Soil, Host-associated human, Human digestive system, Engineered, Marine, Freshwater,

Host-associated mammals, Host-associated plant, Forest soil and Grassland that are used for constructing the phylogenetic tree.

2.2. Methods

To explore better microbial clustering approach, we considered four popular phylogenetic trees named Neighbor-Joining (NJ), UPGMA, Maximum Parsimony (MP), Maximum Likelihood (ML) method. A short description of these four phylogenetic trees is given below.

(i) UPGMA (Unweighted Pair Group Method With Arithmetic Mean) approach:

The **UPGMA** (**U**nweighted **P**air **G**roup **M**ethod with **A**rithmetic Mean) method (Sneath, P.H. and Sokal, R.R., 1973) is a simple agglomerative hierarchical clustering method to produce a dendrogram from a distance matrix. The UPGMA method employs a sequential clustering algorithm, in which local topological relationships are inferred in order of decreasing similarity and a dendrogram is built in a stepwise manner. The method UPGMA is a hierarchical clustering method. From the closest two clusters, it makes a higher-level cluster at each step. It defines the distance between two clusters A and B to be the mean distance between elements of each cluster, which means the average of all the distances between all objects $x \in A$ and $y \in B$:

$$d(A,B) = \frac{1}{|A| + |B|} \sum_{x \in A} \sum_{y \in B} d(x,y)$$

The method uses the "Molecular clock hypothesis" as an evolutionary model, which means that it assumes a constant rate of evolution. In phylogenetic, it is not a widely used approach since it assumes certain relationships between the sequences, without these relationships being tested for the actual data. For this reason, it is only used to produce guide-trees in the process of other sophisticated phylogenetic constructional approaches. It was shown that the optimal time for constructing the UPGMA tree is $O(n^2)$.

(ii) The Neighbor-Joining (NJ) Method

For building phylogenetic trees these methods have become the most popular method. NJ builds a tree from the distance matrix: it contains the pairwise evolutionary distances between the elements of a set of sequences (Saitou and Nei, 1987, Gascuel and Steel, 2006).

When starting, we have a star-like tree which is yet unclear. The algorithm iterates the following steps:

Step-1: First, the algorithm calculates a new matrix *Q* as follows. Based on the distance matrix with n sequences, compute

$$Q(i,j) = (n-2)d(i,j) - \sum_{k=1}^{n} d(i,k) - \sum_{k=1}^{n} d(j,k)$$

Where *d(i,j)* is the distance between the *i*th and *j*th sequences.

Step-1: It finds a pair of *i* and *j (i j)* of sequences for which Q*(i,j)* has the minimum value. Now we create a new node attached to the central node and attach the *i*th and *j*th nodes to it. *f* and *g* denote these nodes, as being attached to the new node *u.*

Step-2: We calculate the distance of the sequences attached to the new node from the other sequences using formula

$$\delta(f,u) = \frac{1}{2}d(f,g) + \frac{1}{2(n-2)}\left[\sum_{k=1}^{n} d(f,k) - \sum_{k=1}^{n} d(g,k)\right]$$

$$\delta(g,u) = \delta(f,g) - \delta(f,u)$$

Where *f* and *g* denote the paired nodes joined to *u*, $\delta(f,u)$ and $\delta(g,u)$ denote the branchlength between its two nodes, and these lengths will not, be affected by the later steps of the construction.

Step-3: Now we calculate the distances of the new node *u*, from the other sequences (which are not *f* and *g).* For each sequence *k* that was not involved in the previous step, let

d(u, k) =½ [d(f, k) + *d(g, k)* - *d(f, g)*]

Step-4: Now we iterate these steps using the newly generated node *u* instead of *f* and *g*, with the recalculated distances

For a set of *n* sequences, the algorithm iterates *(n* —3) times. First, it constructs an *n* x *n* size matrix *Q*, then an *(n* — f) x (n — 1) size one, etc. This implementation leads to time complexity O(n^3), but there exist implementations that, are a lot, faster in average.

(iii) Maximum Likelihood (ML) Approach

To estimate the parameters of a statistical model maximum likelihood estimation methods can be used. In phylogenetics, the maximum likelihood method uses statistical techniques and assigns probabilities for a group of possible phylogenetic trees. It is especially precise for building molecular phylogenies *(*Yang Z, 1993 and Reilly et al. 2016). The desired probability is the product of the probabilities of the branches. i.e.

$$L(\text{tree}) = L_0 L_1 L_2 \ldots$$

and the probabilities of the tree is the sum of the probabilities of the individual trees. i.e.

$$L(\text{tree}) = L(\text{tree1}) + L(\text{tree2}) + L(\text{tree}) + \ldots$$

According to several computer studies, ML has the capability to find a correct tree in a relatively short time and works fine for distantly related data sets. Another strength of the method is that it can compare different trees with different evolutionary models within a statistical framework. However, the drawback is that even with small data sets, it is impossible for the method to find the optimal tree for sure and since it, requires to search all the combinations of tree topology and branch length, it, is a computationally expensive method if used for more than a few sequences. There are several heuristics to fasten the approach, like branch-and-bound or the pruning algorithm, which is a variant of dynamic programming. It was not shown to be NP-complete to search for optimal tree topologies defined by likelihood, but finding an optimal tree is still challenging as a branch and bound search is not yet practical for the accustomed representation of phylogenetic trees.

Algorithm:

By using programming, tree topology and branch lengths can efficiently calculate *Pr(D/T, M)* that is all internal states don't have to calculate.

Finding the greatest maximum likelihood tree is expensive

- ✓ Must consider all topologies
- ✓ Find best edge lengths for each topology

Knowledge: use EM algorithm, to optimize these lengths

Given a probabilistic model for nucleotide (or protein) substitution (e.g., Jukes & Cantor), pick the tree that has the highest probability of generating observed data i.e. Given data *D* and model *M*, first find tree *T* such that *Pr(D/T, M)* is maximized

The probability from nucleotide *i* to *j* in time *t*, the model gives values $p_{ij}(t)$,

Makes assumptions for 2 independence

- ✓ Different sites evolve independently
- ✓ Diverged sequence (or species) evolve independently after diverging

For *i*th site If D_i is data then

$Pr(D/T, M) = \Pi_i Pr(D_i/T, M)$

To calculate $Pr(D_i/T, M)$ using the following folmula:

$P_{xy}(t)$~ prob of going from *x* to *y* in time *t*

$Pr(i, j, k, l/T, M) = \Sigma_x \Sigma_y \Sigma_z Pr(x)(P_{xl}. (t_1+t_2+t_3). P_{xy}(t_1).P_{yk} (t_2+t_3).P_{yz}(t_2).P_{zi}(t_3).P_{zj}(t_3))$

(iv) Maximum Parsimony

Maximum-parsimony (MP) is an optimality criterion under which the phylogenetic tree that minimizes the total number of character-state changes is to be preferred *(*Farris, J.S., 2008). Under the MP criterion, the optimal tree will minimize the amount of homoplasy (i.e., convergent evolution, parallel evolution, and evolutionary reversals). The MP is a character-based method that builds a phylogenetic tree by minimizing the total tree length. It searches for the minimum number of evolutionary steps required to explain a given set of data. These steps are for instance substitutions between DNA sequences. The approach searches for all of the possible tree topologies from the given input, data, and chooses the optimal (minimal) tree. The optimum is usually called the most parsimonious tree. Searching for the optimal tree can be computationally hard because the number of possible trees rapidly grows with the number of input sequences. For example, the number of rooted trees in case of *n* input sequences is

$$N_r = \frac{(2n-3)!}{2e^{(n-2)}(n-2)!}$$

While the number of unrooted trees is

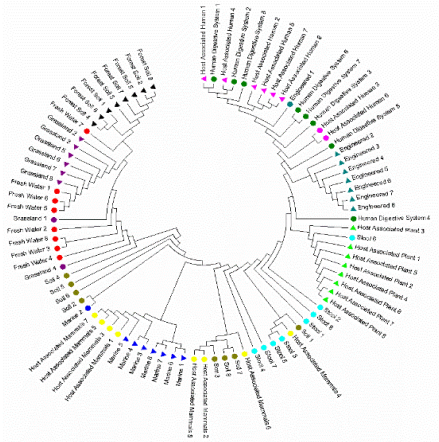
$$N_u = \frac{(2n-5)!}{2e^{(n-3)}(n-3)!}$$

This rapid growth makes it impossible to apply the method on vast data sets. When applied, maximum parsimony describes a non-parametric statistical method for possible relations. However, it, is not consistent, since the probability of producing the evolutionary correct, a tree is not very high under certain conditions.

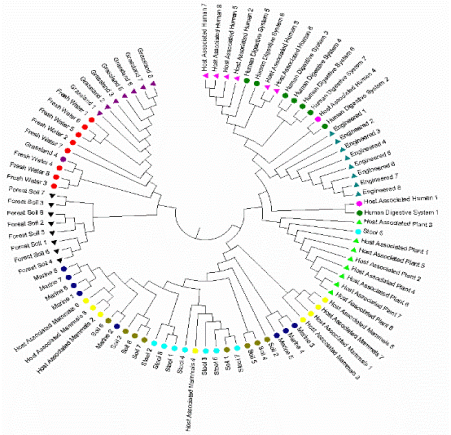
3. Results and Discussions

To investigate the performances of different unsupervised methods for microbial clustering and classification, we analyzed microbial sequence dataset. In Figure 1(a), we see that Neighbor-Joining method correct classification rate of 11 microbial environments as Soil (25%), Host-associated human (75%), Human digestive system (25%), Engineered (87.5%), Marine (87.5%), Freshwater (100%), Host-associated mammals (50%), Host-associated plant (87.5%), Human gut bacteria (37.5%), Forest soil (25%) and Grassland (75%).

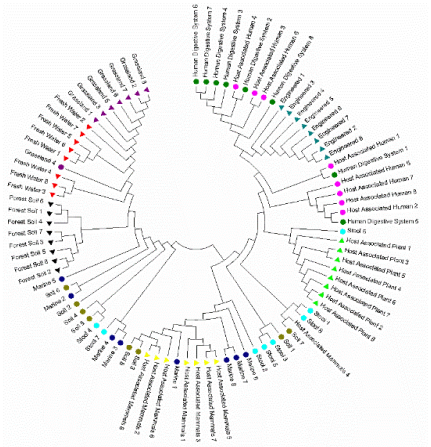
(a) Neighbour-joining approach



(b) UPGMA approach



(c) Maximum parsimony approach



(d) Maximum likelihood approach

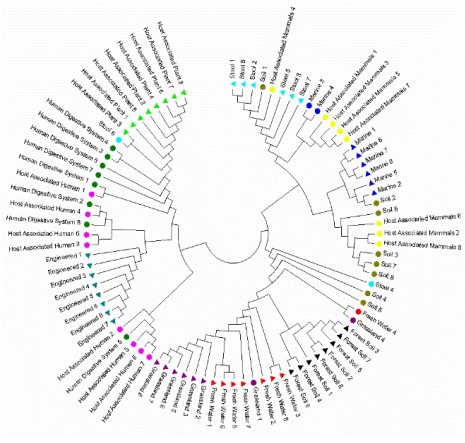


Figure 1: Phylogenetic tree of microbial sequences produced by (a) Neighbor-joining approach (b) UPGMA approach (c) Maximum parsimony approach, and (d) Maximum likelihood approach

That means this method approximately classify 5 microbial environmental sequences named Host-associated human, Engineered, Marine, Host-associated plant and Freshwater considering the cutting threshold 0.6 out of 1.

In Figure 1(b), we see that UPGMA method correct classification rate of 11 microbial environments as Soil (37.5%), Host-associated human (75%), Host-associated mammals (50%), Host-associated plant (100%), Human gut bacteria (37.5%), Forest soil (37.5%) and Grassland (87.5%). That means this method approximately classify 3 microbial environmental sequences named Host-associated human, Host-associated plant and Grassland considering the cutting threshold 0.6 out of 1. In Figure 1(c), we observe that Maximum Parsimony method correct classification rate of 10 microbial environments as Soil (37.5%), Host-associated human (50%), Human digestive system (50%), Engineered (100%), Marine (50%), Freshwater (100%), Host-associated mammals (50%), Host-associated plant (100%), Human gut bacteria(37.5%), Forest soil (37.5%) and Grassland (75%). That means this method approximately classify 4 microbial environmental sequences named Engineered, Freshwater, Grassland and Host-associated plant considering the cutting threshold 0.6 out of 1.In Figure 1(d), we observe that Maximum Likelihood method correct classification rate of 11 microbial environments as Soil (25%), Host-associated human (62.5%), Human digestive system (50%), Engineered (100%), Marine (75%), Freshwater (100%), Host-associated mammals (75%), Host-associated plant (62.5%), Human gut bacteria(75%), Forest soil (37.5%) and Grassland (75%). That means this method approximately classify 8 microbial environmental sequences named Host-associated human, Engineered, Marine, Freshwater, Host-associated mammals, Host-associated plant, Human gut bacteria and Grassland considering the cutting threshold 0.6 out of 1.

Thus, from the above discussion of microbial clustering by using the phylogenetic tree based on sequence dataset we conclude that the maximum likelihood method is comparatively better than the other three methods.

4. Conclusion

The 16S rRNA sequences are now commonly using for identification, classification and quantitation of microbes in complex biological mixtures such as environmental samples like water, soil or air, and microbiome samples, since these sequence are highly conserved between different species of microbes. Unsupervised microbial clustering is now playing a vital role to identify the diseases related virus and bacteria. Phylogenetic tree approach is one of the most popular unsupervised approaches for microbial clustering. In this paper we

performed a comparative study on several approaches including neighbor-joining, UPGMA, maximum parsimony and maximum likelihood methods for phylogenetic tree constraction based on the 16S rRNA sequences collected from 11 environments. The data analysis results shows that maximum likelihood bassed phylogenetic tree outperform the other phylogenetic tree approaches. Thus the output of this paper may help the metagenomic researchers to select a better method of phylogenetic tree construction for microbial clustering to select the diseases related virus more accurately.

References

[1] Doyle, J. J. and Luckow, M. A., (2003). The rest of the iceberg. Legume diversity and evolution in a phylogenetic context. Plant physiology, 131(3), pp.900-910.

[2] Farris, J. S., (2008). Parsimony and explanatory power. Cladistics, 24(5), pp.825-847.

[3] Fitz Gibbon, S. T. and House, C. H., (1999). Whole genome-based phylogenetic analysis of free-living microorganisms. Nucleic acids research, 27(21), pp.4218-4222.

[4] Gascuel, O. and Steel, M., (2006). Neighbor-joining revealed. Molecular biology and evolution, 23(11), pp.1997-2000.

[5] Hall, A. E., Fiebig, A. and Preuss, D., (2002). Beyond the Arabidopsis genome: opportunities for comparative genomics. Plant Physiology, 129(4), pp.1439-1447.

[6] Horner, D. S. and Pesole, G., (2004). Phylogenetic analyses: a brief introduction to methods and their application. Expert Review of Molecular Diagnostics, 4(3), pp.339-350.

[7] O'Reilly, J. E., Puttick, M. N., Parry, L., Tanner, A. R., Tarver, J. E., Fleming, J., Pisani, D. and Donoghue, P. C., J. (2016). Bayesian methods outperform parsimony but at the expense of precision in the estimation of phylogeny from discrete morphological data. Biology Letters, 12, p.20160081.

[8] Patel, J. B., (2001). 16S rRNA gene sequencing for bacterial pathogen identification in the clinical laboratory. Molecular diagnosis, 6(4), pp.313-321.

[9] Saitou, N. and Nei, M., (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular biology and evolution, 4(4), pp.406-425.

[10] Sneath, P. H. and Sokal, R. R., (1973). Numerical taxonomy. The principles and practice of numerical classification.

[11] Soltis, D. E. and Soltis, P. S., (2003). The role of phylogenetics in comparative genetics. Plant Physiology, 132(4), pp.1790-1800.

[12] Tringe, S. G. and Hugenholtz, P., (2008). A renaissance for the pioneering 16S rRNA gene. Current opinion in microbiology, 11(5), pp.442-446.

[13] Weisburg, W. G., Barns, S. M., Pelletier, D. A. and Lane, D. J., (1991). 16S ribosomal DNA amplification for phylogenetic study. Journal of bacteriology, 173(2), pp.697-703.

[14] Woese, C. R., (1987). Bacterial evolution. Microbiological reviews, 51(2), p.221.

[15] Woese, C. R., Stackebrandt, E., Macke, T. J. and Fox, G. E., (1985). A phylogenetic definition of the major eubacterial taxa. Systematic and applied microbiology, 6, pp.143-151.

[16] Yang, Z., (1993). Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Molecular biology and evolution, 10(6), pp.1396-1401.