Heritage and Sustainable Development ISSN 2712-0554
Vol. 1, No. 2, December 2019, pp.78-83

Clustering 16S rRNA for OTU prediction: A similarity based method

Mehmet Can, Osman Gürsoy
Faculty of Engineering and Natural Sciences, International University of Sarajevo, Bosnia

*Corresponding author: mcan@ius.edu.ba

© The Author 2020. Published by ARDA.

Abstract

To study the phylogeny and taxonomy of samples from complex environments Next-generation sequencing (NGS)-based 16S rRNA sequencing, which has been successfully used jointly with the PCR amplification and NGS technology. First step for many downstream analyses is clustering 16S rRNA sequences into operational taxonomic units (OTUs). Heuristic clustering is one of the most widely employed approaches for generating OTUs in which one or more seed sequences to represent each cluster are selected. In this work we chose five random seeds for each cluster from a genes library, and we present a novel distance measure to cluster bacteria in the sample. Artificially created sets of 16S rRNA genes selected from databases are successfully clustered with more than %98 accuracy, sensitivity, and specificity.

Keywords: 16S rRNA; OUT Prediction; Similarity; Sequencing

1 Introduction

Bacteria play an important role in human health and disease [1]. In addition, they have an essential role in various biogeochemical activities. To understand the bacterial world around us, characterizing the taxonomic community composition taken from an environmental sample is very important [2] [3]. Most widely used biomarker for microbial community descriptions is the 16S rRNA (ribosomal RNA) marker genes generated by high-throughput sequencing technology [4]. Advanced sequencing technology can produce millions of 16S rRNA, bypassing the necessity of isolating single organisms for cultivation, and has become a powerful tool for in-depth analysis of bacterial community composition [5], [6].

For rapidly processing the 16S sequencing data, first step is to cluster them into the OTUs [7] [8], which form the basis for estimating the species, diversity, composition, and richness of the microbes in the environment [9][10]. For binning 16S rRNA sequences there are two major approaches:

a. taxonomy dependent methods, where each query sequence is compared against a reference taxonomy database and assigned to the organism of the best-matched annotated sequence using sequence searching [11] or classification [12][13], and taxonomy independent methods (also called de novo clustering) [14], where sequences are grouped into OTUs based on pairwise sequence similarities. However,
b. The success of taxonomy dependent methods are limited by the completeness of reference databases [15] since a significant portion of bacteria in a sample belong to unknown taxa which are not recorded in databases, In contrast, de novo clustering methods divide sequences into OTUs without needing any reference database and have become the preferred choice for researchers [16].

The wide variety of de novo clustering methods has been proposed for binning OTUs in the past decades, can be categorized further into i) hierarchical clustering, ii) heuristic clustering, iii) model-based and iv) network-

based methods [17]. Hierarchical clustering methods like mothur [17], HPC-CLUST [19], ESPRIT [20], and mcClust [21] require a distance matrix. This matrix is computed from all sequences pairs after pairwise sequence alignment or a multiple sequence alignment. Then a hierarchical tree is built, and with a predefined threshold, sequences are assigned into OTUs.

On the other hand, network-based methods like M-pick [22] and DMclust [23] by computing all pairwise sequences distances, first construct a fully connected graph and then by modularity community detection, generates OTUs. Therefore, the computational complexity of both hierarchical and network-based methods is O(N2), where N is the number of sequences [17][23].
Model-based methods, CROP [24] and BEBaC [25] mainly apply some statistical model just like Bayesian model, or a mathematical framework like Gaussian mixture model to describe sequence data. Then based on probability theory, they assign sequences to OTUs. However, they have still a high computational burden [26]. For this reason, hierarchical clustering, model-based and network-based clustering methods, in dealing large-scale sequencing data, quickly meet with the limitations of computational time and memory usage [17].

2 Materials and methods

In this research work, we employ a novel taxonomy dependent method, where each query sequence is compared against reference taxonomy databases in Greengenes, and SILVA, and assigned to the organism of the best-matched. 16S rRNA gene sequences in seven taxonomic classes in Greengenes, and SILVA 16S rRNA libraries are used to create sample sets to be clustered. From each class at a taxonomy level a number of seeds are randomly selected. Using Longest Common Subsequence Search method, the similarity of query sequence with the seed sequences are calculated. If at least one of the similarities with seeds exceeds a certain threshold, the query is assigned the cluster of seeds.

The Longest Common Subsequence Search method helps us to avoid long sequences of pair wise or globally aligned sequences.

2.1 Longest common subsequence search

To find the level of similarity of two gene sequences using Longest Common Subsequence Search method, assume in Figure 1., (a) is a gene reported for a bacteria, and (b) is a gene reported for another, or the same bacteria.

(a) GGCTAACTAGTGTAGAGGTGAAATGATTTAGAT TAGGTGGCAA….

(b) .....GTGTAGAGGTGAAATGCGTAGAT

Figure 1. The longest common subsequence of two genes

The longest common subsequence of (a) and (b) is

GTGTAGAGGTGAAATG

Then we remove this common subsequence from both sequences. Then look for next longest common substring. If there is no longer one this time the string

TAGAT

may be the second longest common subsequence. It is seen that ten iterations of this process is optimal.
Then we add the lengths of these common substrings and normalize by dividing this sum, to the length of the shorter gene.

2.2 Inclass and interclass similarities

The average inclass similarities and interclass averages are compared through the analysis of data contained in the high quality ribosomal RNA databases Greengenes, SILVA, and RDP. The number of non-redundant bacterial 16S ribosomal RNA (rRNA) gene sequences with around 1,200 base pairs is 198.510 for Greengenes. This number is 1.488.662 for SILVA, and 1.350.270 for RDP.

The average inclass similarities and interclass averages are computed for family, genus and species taxon levels in the three databases Greengenes, SILVA, and RDP. The results are show n Tables 2-3.

Table 1. similarities in class/ Inter Class for family level

	Databases	In Class	Inter Class
Phylum	Greengenes	17.47	11.80
	SILVA	29.36	10.23
	RDP	21.86	14.28
	Mean	22.90	12.10
Class	Greengenes	22.64	12.13
	SILVA	21.15	9.63
	RDP	26.47	10.85
	Mean	23.42	10.87
Order	Greengenes	26.57	12.43
	SILVA	33.28	17.55
	RDP	29.99	11.61
	Mean	29.95	13.86
Family	Greengenes	32.54	13.20
	SILVA	56.41	11.45
	RDP	42.40	22.90
	Mean	43.78	15.85
Genus	Greengenes	45.55	13.81
	SILVA	31.50	15.58
	RDP	49.60	16.61
	Mean	42.22	15.33
Species	Greengenes	56.02	10.45
	SILVA	24.23	12.31
	Mean	40.13	12.70
	Overall Mean	30.08	13.45

It is seen that there is a significat difference between in class and inter class similarities for three important taxon levels. Hence this observation shows that longest common sequence similarity measure can be used for both annotation and clustering of unknown samples [27].

3 Results

Three 16S rRNA libraries are used with 198,510 genes Greengenes, with 801,984 genes, RDP, and with 1,820,420 genes SILVA are used to show the accuracy, sensitivity, and specificity of LCSS clustering technique.

At each taxonomic level, 50 genes are selected from each of 20 classes. These 1000 genes are then shuffled. From each class five seeds are randomly selected. Then the Longest Common Subsequence similarities of seeds to a sample gene (query) are calculated. If any of five seeds is similar to the query gene beyond a threshold, this query is put in the same cluster as these seeds.

Using this technique, 1000 genes are clustered with the Accuracy, Sensitivity, and specificity in Table 4 for all taxonomic classes.

Table 2. Accuracy, Sensitivity, and specificity of clustering in Greengenes

%	Accuracy	Sensitivity	specificity
Phylum	98.76	67.90	98.69
Class	97.12	95.35	97.07
Order	97.26	94.96	97.23
Family	97.06	95.21	96.99
Genus	97.75	85.30	98.27

Species	97.00	98.30	96.93

Table 3. Accuracy, Sensitivity, and specificity of clustering in RDP

%	Accuracy	Sensitivity	specificity
Phylum	94.73	65.30	94.45
Class	88.49	62.30	87.88
Order	88.64	76.90	89.25
Family	74.78	83.30	74.33
Genus	88.91	83.30	84.33

Table 4. Accuracy, Sensitivity, and specificity of clustering in SILVA

	Accuracy	Sensitivity	specificity
Phylum	94.64	89.10	94.35
Class	99.12	70.00	98.83
Order	94.22	80.30	94.95
Family	98.06	67.00	99.69
Genus	95.39	96.50	94.74
Species	93.51	64.20	95.05

4 Conclusion

16S rRNA high-throughput sequencing has become a powerful and convenient technology for studying microbial diversity and composition in the environmental samples. Until now, numerous heuristic clustering methods have been developed to pick OTUs, but most of them just select one sequence as the cluster seed, resulting in OTUs overestimation and sensitivity to the sequencing errors. In this work, we proposed a novel similarity clustering method (namely LCSSM).

References

[1] Fuks, G., Elgart, M., Amir, A., Zeisel, A., Turnbaugh, P. J., Soen, Y., et al. (2018). Combining 16S rRNA gene variable regions enables high-resolution microbial community profiling. Microbiome 6:17. doi: 10.1186/s40168-017-0396-x

[2] Lane, D. J., Pace, B., Olsen, G. J., Stahl, D. A., Sogin, M. L., and Pace, N. R. (1985). Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. Proc. Natl. Acad. Sci. U.S.A. 82, 6955–6959.

[3] Wei, Z. G., Zhang, S. W., and Jing, F. (2016). Exploring the interaction patterns among taxa and environments from marine metagenomic data. Quant. Biol. 4, 84–91.

[4] Koslicki, D., Foucart, S., and Rosen, G. (2013). Quikr: a method for rapid reconstruction of bacterial communities via compressive sensing. Bioinformatics 29, 2096–2102. doi: 10.1093/bioinformatics/btt336

[5] Zhang, S. W., Wei, Z. G., Zhou, C., Zhang, Y. C., and Zhang, T. H. (2013). "Exploring the interaction patterns in seasonal marine microbial communities with network analysis," in Proceedings of the International Conference on Systems Biology, Huangshan, 63–68.

[6] Wei, Z.-G., and Zhang, S.-W. (2018). NPBSS: a new PacBio sequencing simulator for generating the continuous long reads with an empirical model. BMC Bioinformatics 19:177. doi: 10.1186/s12859-018-2208-0

[7] Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C., Knight, R., and Gordon, J. I. (2007). The human microbiome project: exploring the microbial part of ourselves in a changing world. Nature 449, 804–810.

[8] Peterson, J., Garges, S., Giovanni, M., Mcinnes, P., Wang, L., Schloss, J. A., et al. (2009). The NIH human microbiome project. Genome Res. 19, 2317–2323. doi: 10.1101/gr.096651.109

[9] Amir, A., Mcdonald, D., Navas-Molina, J. A., Kopylova, E., Morton, J. T., Xu, Z. Z., et al. (2017). Deblur rapidly resolves single-nucleotide community sequence patterns. mSystems 2:e00191-16. doi: 10.1128/mSystems. 00191-16 Frontiers in Microbiology | www.frontiersin.org 11 March 2019 | Volume 10 | Article 428 Wei and Zhang Dynamic Multi-Seeds Clustering Method

[10] Westcott, S. L., and Schloss, P. D. (2017). OptiClust, an improved method for assigning amplicon-based sequence data to operational taxonomic units.mSphere 2:e00073-17. doi: 10.1128/mSphereDirect.00073-17

[11] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

[12] Liu, Z., Pan, Q., Dezert, J., and Martin, A. (2017). Combination of classifiers with optimal weight based on evidential reasoning. IEEE Trans. Fuzzy Syst. 26, 1217–1230.

[13] Liu, Z., Pan, Q., Dezert, J., Han, J.-W., and He, Y. (2018). Classifier fusion with contextual reliability evaluation. IEEE Trans. Cybern. 48, 1605–1618. doi: 10. 1109/TCYB.2017.2710205

[14] Chen, W., Zhang, C. K., Cheng, Y., Zhang, S., and Zhao, H. (2013b). A comparison of methods for clustering 16S rRNA sequences into OTUs. PLoS One 8:e70837. doi: 10.1371/journal.pone.0070837

[15] Chen, S. Y., Deng, F., Huang, Y., Jia, X., Liu, Y. P., and Lai, S. J. (2016). bioOTU: an improved method for simultaneous taxonomic assignments and operational taxonomic units clustering of 16s rRNA gene sequences. J. Comput. Biol. 23, 229–238. doi: 10.1089/cmb.2015.0214

[16] Cai, Y., Wei, Z., Jin, Y., Yang, Y., Mai, V., Qi, M., et al. (2017). ESPRIT-Forest: parallel clustering of massive amplicon sequence data in subquadratic time. PLoS Comput. Biol. 13:e1005518. doi: 10.1371/journal.pcbi.1005518

[17] Wei, Z. G., Zhang, S. W., and Zhang, Y. Z. (2017). DMclust, a density-based modularity method for accurate OTU picking of 16S rRNA sequences. Mol. Inform. 36:1600059. doi: 10.1002/minf.201600059

[18] Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., et al. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microbiol. 75, 7537–7541. doi: 10.1128/AEM. 01541-09

[19] Matias Rodrigues, J. F., and von Mering, C. (2013). HPC-CLUST: distributed hierarchical clustering for large sets of nucleotide sequences. Bioinformatics 30, 287–288. doi: 10.1093/bioinformatics/btt657

[20] Sun, Y., Cai, Y., Liu, L., Yu, F., Farrell, M. L.,Mckendree, W., et al. (2009). ESPRIT: estimating species richness using large collections of 16S rRNA pyrosequences. Nucleic Acids Res. 37:e76. doi: 10.1093/nar/gkp285

[21] Cole, J. R., Wang, Q., Fish, J. A., Chai, B., Mcgarrell, D. M., Sun, Y., et al. (2013). Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res. 42, D633–D642. doi: 10.1093/nar/gkt1244

[22] Wang, X., Yao, J., Sun, Y., andMai, V. (2013). M-pick, a modularity-based method for OTU picking of 16S rRNA sequences. BMC Bioinformatics 14:43. doi: 10. 1186/1471-2105-14-43

[23] Wei, Z.-G., and Zhang, S.-W. (2017). DBH: a de Bruijn graph-based heuristic method for clustering large-scale 16S rRNA sequences into OTUs. J. Theor. Biol. 425, 80–87. doi: 10.1016/j.jtbi.2017.04.019

[24] Hao, X., Jiang, R., and Chen, T. (2011). Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering. Bioinformatics 27, 611–618. doi: 10.1093/bioinformatics/btq725

[25] Cheng, L., Walker, A. W., and Corander, J. (2012). Bayesian estimation of bacterial community composition from 454 sequencing data. Nucleic Acids Res. 40, 5240–5249. doi: 10.1093/nar/gks227

[26] Chen, W., Cheng, Y., Zhang, C., Zhang, S., and Zhao, H. (2013a). MSClust: a multiseeds based clustering algorithm for microbiome profiling using 16S rRNA sequence. J. Microbiol. Methods 94, 347–355. doi: 10.1016/j.mimet.2013.07.004

[27] Can, M., and Gursoy, O. , Taxonomic Classification of Bacteria Using Common Substrings Southeast Europe Journal of Soft Computing Vol.8 No.1 March 2019 (1-4)