VOL9 NO.1 March 2020 - ISSN 2233 – 1859

Operations Research Society in Bosnia and Herzegovina

Southeast Europe Journal of Soft Computing
Available online: http://scjournal.ius.edu.ba

IUS Soft Computing Research Group

A De Novo Clustering Method: Snowball for Assigning 16S rRNA Gene Sequences to Operational Taxonomic Units

M. Can
O. Gürsoy

Faculty of Engineering and Natural Sciences,
International University of Sarajevo International University of Sarajevo,
Hrasnicka Cesta 15, Ilidža 71210 Sarajevo,
Bosnia and Herzegovina
mcan@ius.edu.ba
ogursoy@ius.edu.ba

Article Info

Article history:
Article received on 10 January 2020
Received in revised form 1 February 2020

Keywords:
16S rRNA gene, LongestCommonSubsequence, Taxonomic clustering, Snowball

ABSTRACT: To analyze complex biodiversity in microbial communities, 16S rRNA marker gene sequences are often assigned to operational taxonomic units (OTUs). The abundance of methods that have been used to assign 16S rRNA marker gene sequences into OTUs brings discussions in which one is better. Suggestions on having clustering methods should be stable in which generated OTU assignments do not change as additional sequences are added to the dataset is contradicting some other researches contend that the methods should properly present the distances of sequences is more important. We add one more de novo clustering algorithm, Rolling Snowball to existing ones including the single linkage, complete linkage, average linkage, abundance-based greedy clustering, distance-based greedy clustering, and Swarm and the open and closed-reference methods. We use GreenGenes, RDP, and SILVA 16S rRNA gene databases to show the success of the method. The highest accuracy is obtained with SILVA library.

1. INTRODUCTION

By the throughput of next-generation sequencing technologies, microbial ecologists are able to generate millions of 16S rRNA gene sequences in a reasonable time. Characterizing the composition of millions of sequences from hundreds of samples became very popular. To understand the complexity of biodiversity within the large microbial datasets, sequences are often clustered into meaningful bins commonly known as operational tax¬onomic units (OTUs). To study the biodiversity within samples and between different samples, OTUs are used (Schloss & Westcott, 2011).

Characterization of the biodiversity associated with the Gilbert et al., 2011), and soil (e.g., Shade et al., 2013) provide such comparisons to researchers. Aligned sequences are clustered into OTUs using a threshold of 97% similarity or a distance of 3%, according to a convention emerged within the field of microbial ecology. The definition of the bins is operational since it can be changed to fit the needs of the particular project, and this is one of the advantages of the OTU-based approach.

Software such as mothur (Schloss et al., 2009), QIIME (Caporaso et al., 2010), and other tools (Sun et al., 2009; Edgar, 2010; Edgar, 2013; Cai & Sun, 2011; Mah´e et al., 2014), use several clustering techniques, and it is important

this conventional OTU threshold of 97% similarity or a distance of 3%. It is also necessary to understand how the selected method affects the precision and accuracy of assigning sequences to OTUs. Mainly, three approaches have been developed to assign sequences to OTUs.

1.1. Closed-reference Clustering

Phylotyping (Schloss & Westcott, 2011) or closed-reference clustering (Navas-Molina et al., 2013) method compares sequences to a curated reference database and clusters them into the same OTU that are similar to the same reference sequence. (Can, and Gürsoy, 2019a,b,c, Gursoy, and Can, 2019) When the reference does not adequately reflect the biodiversity of the community, reference-based clustering methods trouble. New sequences that are not in the reference database cannot be assigned to any OTU. It is also reported that the commonly used hypervariable regions within the 16S rRNA marker genes do not evolve at the same rate as the full-length gene (Schloss, 2010; Kim, Morrison & Yu, 2011). Hence, a sequence representing a small fragment of the gene could be more than 97% similar to many other reference sequences. Moreover, although two sequences might be 97% similar to the same reference sequence they may only be 94% similar to each other. Constructing OTUs with the closed-reference approach is problematic because of this as well.

It is also possible that a sequence may be equally similar to two or more reference sequences. An alternative way to closed-reference approach is to employ a classifier to assign taxonomy to each sequence so that clustering becomes possible at the desired level within the Linnean taxonomic hierarchy (Schloss & Westcott, 2011). Reference-based methods are superior because of their speed, ability to compare OTU assignments across studies, potential for trivial parallelization, and the hope that as databases improve, the OTU assignments will also improve.

1.2. de novo Clustering

Distance-based (Schloss & Westcott, 2011) or de novo clustering (Navas-Molina et al., 2013) uses the distance between sequences rather than a reference database to cluster sequences into OTUs. In comparison to the efficiency of the closed-reference clustering method, computational cost increases quadratically with the number of unique sequences in a hierarchical de novo clustering method. Furthermore, sequencing errors also cause inflation of the number of unique sequences requiring large amounts of memory and time for clustering. Using stringent quality control measures, error rates can be reduced, then these problems can be cured (Kozich et al., 2013).

To approximate the clustering of hierarchical methods, heuristics are developed to approximate the clustering of hierarchical methods (Sun et al., 2009; Edgar, 2010; Mah´e et al., 2014). Distance-based greedy clustering (DGC), and abundance-based greedy clustering (AGC) (Edgar, 2010; He et al., 2015) are the two related heuristic methods which are implemented in USEARCH described recently. These greedy methods cluster sequences within a defined similarity threshold of an index sequence or create a new index sequence. If a sequence is more similar than the defined threshold, it is assigned to the closest centroid based (i.e., DGC) or the most abundant centroid (i.e., AGC). OTU assignments are sensitive to the input order of the sequences in de novo approaches (Mah´e et al., 2014; He et al., 2015). It is doubtful whether the differences in assignments are meaningful or not. The strength of de novo clustering is its independence of references for carrying out the clustering step. Thus, de novo clustering is popular across the field. After clustering, the classification of each sequence can be used to obtain a consensus classification for the OTU (Schloss & Westcott, 2011).

1.3. Open-reference Clustering

Another approach combining the closed-reference and de novo approaches (Navas-Molina et al., 2013; Rideout et al., 2014) is open-reference clustering. It performs closed-reference clustering followed by de novo clustering for the sequences that are not found in the reference database. One may expect that this method should have the strengths of both closed-reference and de novo clustering but different OTU definitions employed by closed-reference and de novo clustering implementations pose a possible problem.
Classifying sequences to a bacterial family or genus and then assign those sequences to OTUs within those taxonomic groups using the average linkage method (Schloss & Westcott, 2011) is an alternative to this approach. For example, all sequences classified as Porphyromonadaceae would be assigned to OTUs with the average linkage method using a 3% distance threshold. Those sequences that did not classify to a known family would also be clustered using the average linkage method. This approach lends itself nicely to parallelization since each taxonomic group is seen as being independent and can be processed separately, this is an advantage of this technique. Such an approach would overcome the difficulty of mixing OTU definitions between the closed-reference and de novo approaches; Certainly, it will still suffer from the problems associated with database quality and classification error.

1.4. Quality of OTU Assignments

The three broad approaches in the above created many options for assigning sequences. It has been difficult to objectively assess. The quality of OTU assignments is difficult to assess objectively. Some of the assessments are focused on the time and memory required (Sun et al., 2009; Cai & Sun, 2011; Mah´e et al., 2014; Rideout et al., 2014). When judging a clustering method, these are valid parameters, but the quality of the OTU assignments is something else. Some other quality assessment methods use its ability to generate data that parallels classification data (White et al., 2010; Sun et al., 2011; Cai & Sun, 2011). Since bacterial taxonomy often reflects historical biases

amongst bacterial systematicists, hence this approach is problematic. It is well known that the rates of evolution across lineages are not the same (Wang et al., 2007; Schloss, 2010).

Clustering of mock community data to evaluate methods is a common approach (Huse et al., 2010; Barriuso, Valverde & Mellado, 2011; Bonder et al., 2012; Chen et al., 2013; Edgar, 2013; Mah´e et al., 2014; May et al., 2014). All these approaches ignore the effects of sequencing errors that tend to accumulate with sequencing depth. Therefore they are highly idealized communities that lack the phylogenetic diversity of real microbial communities (Schloss, Gevers & Westcott, 2011; Kozich et al., 2013). Other quality assessment techniques measure the quality based on the method's ability to generate the same OTUs as generated by other methods (Rideout et al., 2014; Schmidt, Rodrigues & Mering, 2014b). But it does not solve the fundamental question of which method is optimal.

Sequences that are clustered into the same OTU are expected to have similar ecological affiliations (Koeppel & Wu, 2013; Preheim et al., 2013; Schmidt, Rodrigues & Mering, 2014a). The ecological consistency concept as a metric of quality is an interesting approach and is a quantitative metric. It is unclear how the metric would be objectively validated. Westcott and Schloss (2015) proposed a method for evaluation of OTU assignments using the distances between pairs of sequences (Schloss & Westcott, 2011).

Stability is defined in a rather recent analysis by He and colleagues (2015) as the ability of a method to provide the same clustering on a subset of the data as was found in the full dataset. They characterized the three general clustering approaches by focusing on what they called stability. Their concept of stability focused on the precision of the assignments. They put the quality of the OTU assignments in second place.

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 (LCS), 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 (LCS) 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. (Can, and Gürsoy, 2019a).

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 in class similarities and interclass averages are computed for all taxon levels in the three databases Greengenes, SILVA, and RDP. The results are shown in Table 1.

It is seen that there is a significat difference between average in class and inter class similarities for all taxon levels. Hence this observation shows that longest common sequence similarity measure can be used for both annotation and clustering of unknown samples [27]. (Can, and Gürsoy, 2019c).

2.3. LCS Similarity Measure is Successfully Used for Annotation

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 LCS clustering technique.

Table 1. In Class/ Inter Class similarities for all taxon levels

	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	30.08	13.45

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 (LCS) 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.

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	94.73	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

Using this technique, 1000 genes are clustered with the Accuracy, Sensitivity, and specificity in Tables 2-4 for all taxonomic classes. (Can, and Gürsoy, 2019).

2. 4. Methods

After obtaining the 16S rRNA gene reads from the high quality ribosomal RNA databases Greengenes, SILVA, and RDP, we have selected 10, 25, 50, 100, and 200 genum classes from each database. From each class we have randomly selected 10, 25, and 50 genes as samples.

2.4.1 Similarity Matrices and Snowball

The selected samples are shuffled to get gene sets with 100, 2500, and 5000 elements. These genes are given a number. Then 100x100, 2500x2500, and 5000x5000 pair wise LCS similarity matrices are calculated. Each row of this nxn matrix consists of similarity levels of n bacteria to the bacteria of this row. Then we sort the n-1 bacteria according to the similarities to this bacterium. We use an optimal threshold to truncate these sorted lists of bacteria numbers.

An Example

From ribosomal RNA database SILVA, 10 genum class, and from each class 10 gene sequences are randomly selected.

Table 5. Labels of bacteria from each of ten genum classes

Gen										
1	1	2	7	10	16	21	29	62	63	97
2	9	36	49	53	68	69	70	71	73	85
3	4	37	38	44	52	57	74	76	84	98
4	5	11	31	32	40	42	80	88	90	93
5	18	28	33	66	75	77	78	79	86	95
6	13	14	23	24	46	51	56	60	94	100
7	19	22	25	45	47	82	83	91	92	99
8	12	27	39	41	58	59	61	65	81	96
9	3	8	17	30	34	35	43	55	64	87
10	6	15	20	26	48	50	54	67	72	89

Then they are shuffled to get a set of 100 sequences. Then 100x100 pairwise similarity matrix is calculated. A sample 10x10 portion of it shown in Table 6.

Table 6. A sample 10x10 portion of the similarity matrix

1.00	0.87	0.19	0.11	0.10	0.08	0.59	0.20	0.20	0.70
0.87	1.00	0.19	0.12	0.10	0.09	0.56	0.22	0.22	0.59
0.19	0.19	1.00	0.14	0.13	0.08	0.19	0.23	0.22	0.19
0.11	0.12	0.14	1.00	0.20	0.09	0.12	0.14	0.14	0.11
0.10	0.10	0.13	0.20	1.00	0.09	0.10	0.11	0.11	0.10
0.08	0.09	0.08	0.09	0.09	1.00	0.09	0.09	0.10	0.08
0.59	0.56	0.19	0.12	0.10	0.09	1.00	0.22	0.22	0.55
0.20	0.22	0.23	0.14	0.11	0.09	0.22	1.00	0.23	0.20
0.20	0.22	0.22	0.14	0.11	0.10	0.22	0.23	1.00	0.20
0.70	0.59	0.19	0.11	0.10	0.08	0.55	0.20	0.20	1.00

Then each row is sorted, and truncated according to an optimal threshold. When similarities are replaced by the numbers of the genes which has this similarity with the query genes, we get the close relatives of the query genes as seen partly in Table 7.

Table 7. Close relatives of the query genes in the first column

1	**63**	**2**	**29**	**21**	**97**	**10**	**16**	**7**	**62**
2	**1**	**29**	**63**	**97**	**21**	**10**	**7**	**16**	**62**
3	30	34							
4	37	57	38	74	98	84	52	44	76
5	88	93	32	11	40	42	90	31	80
6	26	15	50						
7	**62**	**29**	**1**	**97**	**2**	**21**	**10**	**63**	**16**
8	17								
9	71	53	68	73	70	36	69	85	49
10	**1**	**21**	**97**	**63**	**29**	**2**	**16**	**7**	**62**

Rolling Snowball

In Table 7. A hundred bacteria is listed with their close relatives. To obtain clusters from these lists of close relatives, we start by one of the rows. If a row has at least one common element with another row, the union of these two rows is kept and the two rows are erased from the list. For example in Table 7. the rows $\{1, 2, 7, 8, 10\}$ have common elements, and rolling snowball collects them together in the union

$$\{1,2,7,8,10, 16, 18, 21,29,62,63,97\}.$$

search is continued to other rows that has at least one common element with this set, and the rolled rows are erased from the list. At the end of this process the clusters emerge as in Table 8.

Table 8. Clusters emerged after snowball rolling.

1	2	7	10	16	21	29	62	63	97	10
9	36	49	53	68	69	70	71	73	85	10
4	37	38	44	52	57	74	76	84	98	10
5	11	31	32	40	42	80	88	90	93	10
33	54									1
13	14	23	24	46	51	56	60	94	100	10
82	83									2
12	27	39	41	58	59	61	65	81	96	10
3	8	17	30	34	43	55	64			8
6	15	26	50							6

With another list of singletons

$\{18\},\{19\},\{22\},\{23\},\{25\},\{28\},\{33\},\{35\},\{45\},\{47\},\{48\},$

$\{66\},\{67\},\{72\},\{75\},\{77\},\{78\},\{79\},\{83\},\{85\},\{87\},\{89\}$

$,\{91\},\{92\},\{95\},\{99\}$

If we compare the clusters in Table 8, and the labels in Table 5., we get the last column of correct clustering numbers. Accuracy, sensitivity, and specificity are calculated by the following formulas where

TP	True positive
FP	False positive
TN	True negative
FN	False negative
N	Sample size
M	Group size

$$\text{accuracy} = (\text{TP} + \text{TN})/\text{N}$$
$$\text{sensitivity} = \text{TP}/\text{M}$$
$$\text{specificity} = \text{TN}/(\text{N} - \text{M})$$

For this example accuracy=98%, sensitivity=75%, and specificity=99.9%. The singletons are due to the miss clustering.

The above example is repeated ten times. Accuracy, sensitivity, and specificity of re clustering of the sample data is successfully done as in the Table 9.

Table 9. Average accuracy, sensitivity, and specificity of snowball rolling

	RS	TP	FP	TN	FN	acc	sen	spe
1	10	10	0	90	0	1.0	1.0	1.0
2	10	10	0	90	0	1.0	1.0	1.0
3	2	2	0	90	8	0.9	0.2	1.0
4	10	9	1	89	1	0.9	0.9	0.9
5	18	10	8	82	0	0.9	1.0	0.9
6	10	10	0	90	0	1.0	1.0	1.0
7	10	10	0	90	0	1.0	1.0	1.0
8	10	10	0	90	0	1.0	1.0	1.0
9	3	3	0	90	7	0.9	0.3	1.0
10	5	5	0	90	5	0.9	0.5	1.0
Average						0.9	0.8	1.0

3. SNOWBALL FOR SAMPLES FROM GREEN GENES, RDP AND SILVA GENE LIBRARIES

Greengenes 16S rRNA library contains 198,510 genes, while RDP has 801,984 genes, and SILVA 1,820,420 genes. In Genus taxonomy level, from several numbers of classes, several samples are randomly chosen with varying sample sizes. In Table 10., it is shown sampling method, and corresponding average accuracy, sensitivity, and specificities of snowball rolling technique.

For the Greengenes 16S rRNA library, that contains 198,510 genes, in the first experiment, 25 genes are randomly selected from 200 genus classes. In the second, 50 genes are randomly selected from 50 genus classes, in the third, 40 by 40, and in the fourth, 20 by 20. Sampling style, LCS similarity threshold used, number of clusters correctly identified, accuracy, sensitivity, and specificity of Snowball clustering is shown in Table 10.

Table 10. Accuracy, Sensitivity, and specificity of Snowball clustering in Greengenes

Samp	Simil	Cluste	Accura	Sensitiv	specifici
25x200	0.35	145	0.9964	0.482	0.999
50x50	0.40	41	0.9896	0.633	0.998
40x40	0.40	30	0.9881	0.615	0.998
20x20	0.35	17	0.9033	0.705	0.992

As seen in Table 10, accuracy increases by the class size while sensitivity which is the average number of correct clustering decreases, but all of them are at acceptable levels.

For the RDP 16S rRNA library, that contains 801,984 genes, in the first experiment, 25 genes are randomly selected from 200 genus classes. In the second, 50 genes are randomly selected from 100 genus classes, in the third, 50 by 50, and in the fourth, 50 by 20. Sampling style, LCS similarity threshold used, number of clusters correctly identified, accuracy, sensitivity, and specificity of Snowball clustering is shown in Table 11.

Table 11. Accuracy, Sensitivity, and specificity of Snowball clustering in RDP

Sampli	Simi	Clu	Accur	Sensitiv	specificity
25x200	0.85	133	0.99	0.22	0.99
50x100	0.80	75	0.99	0.36	0.99
50x50	0.70	36	0.99	0.42	0.99
50x20	0.70	17	0.96	0.36	0.99

As seen in Table 11, accuracy increases by the class size while sensitivity which is the average number of correct clustering decreases, but all of them are at acceptable levels.

For the SILVA 16S rRNA library, that contains 1,820,420 genes, in the first experiment, 25 genes are randomly selected from 200 genus classes. In the second, 50 genes are randomly selected from 100 genus classes, in the third, 50 by 50, and in the fourth, 10 by 10. Sampling style, LCS similarity threshold used, number of clusters correctly identified, accuracy, sensitivity, and specificity of Snowball clustering is shown in Table 12.

Table 12. Accuracy, Sensitivity, and specificity of Snowball clustering in SILVA

Samplin	Simil	Clust	Accura	Sensitiv	specific
25x200	0.40	169	0.9969	0.5168	0.9993
50x100	0.35	81	0.9940	0.5528	0.9985
50x50	0.55	49	0.9928	0.6576	0.9996
10x10	0.30	9	0.9643	0.7000	0.9963

As seen in Table 12, accuracy increases by the class size while sensitivity which is the average number of correct clustering decreases, but all of them are at acceptable levels.

4. CONCLUSIONS

Clustering algorithms will continue to be developed, as the throughput of next generation sequencing technologies will continue to be improved. Because of timely expansive similarity matrix construction, *De novo* clustering methods are considerably slower and more computationally intensive than reference-based methods. But the greedy *de novo* methods are faster than the hierarchical methods. Removing sequencing error and chimeras is a detriment to execution speed of the *de novo* methods (Kozich et al., 2013). As the rate of sequencing error increases so do the number of unique sequences that must be clustered. The speed of the *de novo* methods requires a four-fold execution time increase when doubling the number of sequences which shows that the scaling is approximately quadratically. Microbial ecologists must continue to refine clustering methods to better handle the size of their growing datasets, but they must also take steps to improve the quality of the underlying data. Ultimately, objective standards must be applied to assess the quality of the data and the quality of OTU clustering.

In this research from Genus taxonomy level, from several numbers of classes, several samples are randomly chosen with varying sample sizes. It is shown that accuracy is related to sampling methods. Average accuracy decreases with the number of clusters in snowball technique.

REFERENCES

Barriuso J, Valverde JR, Mellado RP. (2011) Estimation of bacterial diversity using next generation sequencing of 16S rDNA: a comparison of different workflows. BMC Bioinformatics

12:473 DOI 10.1186/1471-2105-12-473.

Bonder MJ, Abeln S, Zaura E, Brandt, BW. (2012) Comparing clustering and pre-processing in taxonomy analysis. Bioinformatics 28:2891–2897 DOI 10.1093/bioinformatics/bts552.

Cai Y, Sun Y. 2011. ESPRIT-tree: hierarchical clustering analysis of millions of 16S rRNA pyrosequences in quasilinear computational time. Nucleic Acids Research 39:e95–e95 DOI 10.1093/nar/gkr349.

Can, M., and Gürsoy, O., (2019a) Clustering 16S rRNA for OTU prediction: A similarity based method, Heritage and Sustainable Development Vol. 1, No. 2, December 2019, pp.78-83

Can, M. (2019b) Annotation of Bacteria by Greengenes Classifier Using 16S rRNA Gene Hyper Variable Regions, Southeast Europe Journal of Soft Computing Vol.8 No.2 September 2019 (69-78)

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

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pe˜na AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R. (2010) QIIME allows analysis of high-throughput community sequencing data. Nature Methods 7:335–336 DOI 10.1038/nmeth.f.303.

Chen W, Zhang CK, Cheng Y, Zhang S, Zhao H. (2013) A comparison of methods for clustering 16S rRNA sequences into OTUs. PLoS ONE 8:e70837 DOI 10.1371/journal.pone.0070837.

Eddelbuettel D. (2013) Seamless R and C++ integration with Rcpp. New York: Springer.

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460–2461 DOI 10.1093/bioinformatics/btq461.

Edgar RC. (2013) UPARSE: Highly accurate OTU sequences from microbial amplicon reads. Nature Methods 10:996–998 DOI 10.1038/nmeth.2604.

Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. (2011) UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 27:2194–2200 DOI 10.1093/bioinformatics/btr381.

Gilbert JA, Steele JA, Caporaso JG, Steinbr¨uck L, Reeder J, Temperton B, Huse S, McHardy AC, Knight R, Joint I, Somerfield P, Fuhrman JA, Field D. (2011) Defining seasonal marine microbial community dynamics. The ISME Journal 6:298–308 DOI 10.1038/ismej.2011.107.

Gursoy, O., Can, M. (2019) Hypervariable Regions in 16S rRNA Genes for the Taxonomic Classification Southeast Europe Journal of Soft Computing Vol.8 No.1 March 2019 (23-26)

He Y, Caporaso JG, Jiang X-T, Sheng H-F, Huse SM, Rideout JR, Edgar RC, Kopylova E, Walters WA, Knight R, Zhou H-W. (2015) Stability of operational taxonomic units: an important but neglected property for analyzing microbial diversity. Microbiome 3:20 DOI 10.1186/s40168-015-0081-x.

Huse SM, Welch DM, Morrison HG, Sogin ML. (2010) Ironing out the wrinkles in the rare

biosphere through improved OTU clustering. Environmental Microbiology 12:1889–1898 DOI 10.1111/j.1462-2920.2010.02193.x.

Kim M, Morrison M, Yu Z. (2011) Evaluation of different partial 16S rRNA gene sequence regions for phylogenetic analysis of microbiomes. Journal of Microbiological Methods 84:81–87 DOI 10.1016/j.mimet.2010.10.020.

Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. (2013) Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. Applied and Environmental Microbiology 79:5112–5120 DOI 10.1128/AEM.01043-13.

Mah´e F, Rognes T, Quince C, De Vargas C, Dunthorn M. (2014) Swarm: robust and fast clustering method for amplicon-based studies. PeerJ 2:e593 DOI 10.7717/peerj.593.

Matthews B. (1975) Comparison of the predicted and observed secondary structure of t4 phage lysozyme. Biochimica et Biophysica Acta (BBA)Protein Structure 405:442–451 DOI 10.1016/0005-2795(75)90109-9

May A, Abeln S, Crielaard W, Heringa J, Brandt BW. (2014) Unraveling the outcome of 16S rDNA-based taxonomy analysis through mock data and simulations. Bioinformatics 30:1530–1538 DOI 10.1093/bioinformatics/btu085.

Navas-Molina JA, Peralta-S´alez A, McMurdie PJ, V´anchez JM, Gonz´azquez-Baeza Y, Xu Z, Ursell LK, Lauber C, Zhou H, Song SJ, Huntley J, Ackermann GL, Berg-Lyons D, Holmes S, Caporaso JG, Knight R. (2013) Advancing our understanding of the human microbiome using QIIME. In: Methods in enzymology. Amsterdam: Elsevier BV, 371–444.

Rideout JR, He Y, Navas-Molina JA, Walters WA, Ursell LK, Gibbons SM, Chase J, McDonald D, Gonzalez A, Robbins-Pianka A, Clemente JC, Gilbert JA, Huse SM, Zhou H-W, Knight R, Caporaso JG. (2014) Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences. PeerJ 2:e545 DOI 10.7717/peerj.545.

Schloss PD. (2010) The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies. PLoS Computational Biology 6:e1000844 DOI 10.1371/journal.pcbi.1000844.

Schloss PD, Gevers D, Westcott SL. (2011) Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. PLoS ONE 6:e27310 DOI 10.1371/journal.pone.0027310.

Schloss PD, Westcott SL. (2011) Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis. Applied and Environmental Microbiology 77:3219–3226 DOI 10.1128/AEM.02810-10.

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Horn DJV, Weber CF. (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied and Environmental Microbiology 75:7537–7541 DOI 10.1128/AEM.01541-09.

Schmidt TSB, Rodrigues JFM, Von Mering C. (2014a) Ecological consistency of SSU rRNA-based operational taxonomic units at a global scale. PLoS Computational Biology 10:e1003594 DOI 10.1371/journal.pcbi.1003594.

Schmidt TSB, Rodrigues JFM, Von Mering C. (2014b) Limits to robustness and reproducibility in the demarcation of operational taxonomic units. Environmental Microbiology 17:1689–1706 DOI 10.1111/1462-2920.12610.

Shade A, Klimowicz AK, Spear RN, Linske M, Donato JJ, Hogan CS, McManus PS, Handelsman J. (2013) Streptomycin application has no detectable effect on bacterial community structure in apple orchard soil. Applied and Environmental Microbiology 79:6617–6625 DOI 10.1128/AEM.02017-13.

Sun Y, Cai Y, Liu L, Yu F, Farrell ML, McKendree W, Farmerie W. (2009) ESPRIT: estimating species richness using large collections of 16S rRNA pyrosequences. Nucleic Acids Research 37:e76–e76 DOI 10.1093/nar/gkp285.

Sun Y, Cai Y, Huse SM, Knight R, Farmerie WG, Wang X, Mai V. (2011) A large-scale benchmark study of existing algorithms for taxonomy-independent microbial community analysis. Briefings in Bioinformatics 13:107–121 DOI 10.1093/bib/bbr009.

Wang Q, Garrity GM, Tiedje JM, Cole JR. (2007) Naive bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied and Environmental Microbiology 73:5261–5267 DOI 10.1128/AEM.00062-07.

Westcott and Schloss (2015), De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. PeerJ 3:e1487; DOI 10.7717/peerj.1487

White JR, Navlakha S, Nagarajan N, Ghodsi M-R, Kingsford C, Pop M. (2010) Alignment and clustering of phylogenetic markers—implications for microbial diversity studies. BMC Bioinformatics 11:152 DOI 10.1186/1471-2105-11-152.