A De Novo Robust Clustering Approach for Amplicon-Based Sequence Data

Alexandre Bazin Didier Debroas Engelbert Mephu Nguifo

contact@alexandrebazin.com

1 Introduction

Studying the structure of the communities in an ecosystem is central in environmental microbiology [8, 14]. The biosphere's diversity can be determined by amplifying and sequencing specific phylogenetic markers (e.g. 16S rRNA). From there, these amplicons need to be clusterized in "species" named Operational Taxonomic Units (OTUs) [4, 9, 11, 15]. As the volume of sequences has drastically increased in recent times, new clustering tools have emerged to treat the data in reasonable time. The currently used algorithms are, from the point of view of algorithmic complexity, the fastest available that do not produce random results. However, due to their simplicity, the reliability of the results are often discussed. These tools being essentially black boxes, their sensitivity to the sequence order, clustering threshold and structure of the data makes it that the users have no way of knowing whether better Operational Taxonomic Units (OTUs) could have been obtained with different parameters or even whether they correctly represent the data. In these circumstances, there is no choice but to blindly trust them.

Distance-based greedy clustering algorithms such as the ones implemented in OTUclust [1], VSEARCH [13], CD-HIT [10] or USEARCH [5] all share the same base naive algorithm. While more sophisticated algorithms [3, 6, 12, 7, 2] could produce better results quality-wise, their runtime would render them unusable on millions of sequences. As the quality of the OTUs is important, we have to find a way to improve it without increasing the runtime. The different available implementations use a variety of heuristics to counterbalance the simplicity of the algorithm but, to the best of our knowledge, no approach has tried to add a measure of uncertainty to the process. This is why, in order to help increase the quality and trustworthiness of the clustering, we propose to add uncertainty to this simple algorithm through the use of fuzzy clustering.

2 Adding uncertainty to clustering

Distance-based greedy clustering algorithms, such as the one in VSEARCH, produce a number of OTUs and assign each sequence to one of them. The OTU to which a sequence is said to belong to is usually the first one to be encountered that is sufficiently close, i.e. within the specified threshold. This makes a sequence belong only to a single OTU and OTUs either completely include or exclude a sequence. While these would not be problems were the clustering optimal, the need for fast algorithms gives rise to results that are not always trustworthy. The OTUs being presented as absolute, the end user has no choice, should consider them correct and cannot know whether the algorithm has encountered ambiguity. We believe that being less strict in the way the OTUs partition sequences would help produce better results from the end user's point of view. To help increase the quality of the clustering and maximize the information that can be gathered from the data, we propose to add uncertainty to the clustering by means of fuzzy sets.

Using fuzzy OTUs allows us to discern the difference between sequences close to the OTU and sequences extremely far. Using the parameters t_1 and t_2, we can tune the "detection radius" around OTUs to gather information that would normally be discarded by the clustering algorithm.

3 Evaluating fuzzy OTUs

An ideal OTU would contain only sequences with a membership value of 1, meaning a group of sequences has been perfectly regrouped with a good threshold and no sequence lies ambiguously on the border. More realistically, a good OTU would contain many sequences with high membership values and little sequences with low values. A bad OTU with the majority of its sequences having low membership values could mean that the algorithm has chosen as a center a sequence on the border of a group or, even worse, between two distinct groups.

We can quickly evaluate the quality of an OTU with this repartition. If we suppose that each sequence lowers the quality of the OTU depending on its membership value, we can use the following formula :

$$Quality(OTU) = 1 - \sum_{i=1}^{9} \omega_i \times \frac{\text{\# sequences with membership value } i \times 0.1}{\text{\# sequences in the OTU}}$$

with ω_i being the "cost" of having a sequence with membership value $i \times 0.1$.

A problem arises with singletons that always have perfect quality but these can safely be treated separately.

A sequence can belong to multiple OTUs due to fuzzy membership. However, in the end, we want each sequence to be assigned to a single OTU. Hence, we have to choose one of the possible OTUs. We have two types of values left from the clustering process : membership and quality. The first one is based on the distance between the OTU and the sequence and the second one is used to recognize bad OTUs. Choosing the OTU with the best membership value is akin to running VSEARCH. Choosing the OTU with the best quality tends to create bigger OTUs that absorb distant sequences. To better compromise, we can use a linear combination of both values :

$$\alpha \times quality + \beta \times membership$$

Increasing the importance of the quality reduces the number of OTUs containing sequences. When α is low, the "best" OTUs quality-wise absorb very close sequences that would have been attributed to other OTUs. When α gets too high, the best OTUs start absorbing all the sequences around them, effectively acting like an increase of the distance threshold.

4 Experimental Results

4.1 Data

We used our algorithm on a dataset containing 5977 sequences of length between 900 and 3081 for an average of 1442 and taxonomies extracted from the SILVA database. We used a threshold of 0.97 (97% similarity) for determining new OTUs and a threshold of 0.95 for fuzzy membership. For the choice of the OTU for each sequence, we present the results of three strategies : best quality ($\alpha = 1$ and $\beta = 0$), compromise ($\alpha = 0.5$ and $\beta = 0.5$) and distance ($\alpha = 0$ and $\beta = 1$). The comparison with VSEARCH is done using identical parameters when applicable.

The program, dataset and corresponding taxonomy are available on `http://projets.isima.fr/sclust/Expe.html`.

4.2 Relevant Metrics

To measure the effects of introducing uncertainty to the clustering, we consider the computation time, memory usage, number of OTUs, singletons and pairs and average distance in the taxonomy tree between sequences in the same cluster.

4.3 Analysis

Results show that the choice strategy affects every metric relevant to the quality of the clustering : number of OTUs, singletons and pairs, average misclassification. The fuzzy approach uses slightly more memory than VSEARCH but all choice strategies are similar on this metric. When using the default *–maxaccepts* and *–maxrejects* values, computation time is lower for VSEARCH. However, when using higher values for these parameters – and thus more precise clustering - the computation time is the same for both approaches. We observe that increasing the importance of the quality in the OTU choice strategy lowers the final number of OTUs. This is due to the fact that some OTUs are initially created centered on isolated sequences near good OTUs. That isolation lowers their quality and the good OTUs absorb their sequences.

Using the quality also lowers the number of singletons and increases the number of pairs. This most likely means that singletons were created close to either good clusters or one another. The fuzzy approach allows the algorithm to merge those sequences that were slightly too far from the center with their corresponding OTU. The increase in the number of pairs appears to be due to the merging of singletons lying too close to one another. The average taxonomy distance in OTUs is shown to vary wildly. Using only the quality to choose OTUs increases this number as the "best" OTUs attract all the sequences in their fuzzy surroundings. This causes some sequences belonging to different species to be classified together. However, using a compromise between quality and distance lowers this metric as the best clusters only absorb sequences that are sufficiently close to them and should probably be together while rejecting the sequences that are too different.

acknowledgements

This work was supported by the European Union's "*Fonds Européen de Développement Régional (FEDER)*" program and the Auvergne-Rhone-Alpes region.

Method	Time (min)	Memory	#OTUs	#Singletons	#Doubletons	Distance
Fuzzy (best quality)	1:06	652744	3461	2581	442	0.75
Fuzzy (compromise)	1:06	651980	3596	2776	413	0.54
Fuzzy (distance)	1:06	683772	3631	2837	395	0.59
VSEARCH	0:21	632832	3716	2935	388	0.57

Table 1: Results of the clustering.

References

[1] Davide Albanese, Paolo Fontana, Carlotta De Filippo, Duccio Cavalieri, and Claudio Donati. Micca: a complete and accurate software for taxonomic profiling of metagenomic data. *Scientific reports*, 5:9743, 2015.

[2] Violaine Antoine, Benjamin Quost, Marie-Hélène Masson, and Thierry Denoeux. CECM: constrained evidential c-means algorithm. *Computational Statistics & Data Analysis*, 56(4):894–914, 2012.

[3] Violaine Antoine, Benjamin Quost, Marie-Hélène Masson, and Thierry Denoeux. CEVCLUS: evidential clustering with instance-level constraints for relational data. *Soft Comput.*, 18(7):1321–1335, 2014.

[4] Wei Chen, Clarence K Zhang, Yongmei Cheng, Shaowu Zhang, and Hongyu Zhao. A comparison of methods for clustering 16s rrna sequences into otus. *PloS one*, 8(8):e70837, 2013.

[5] Robert C Edgar. Search and clustering orders of magnitude faster than blast. *Bioinformatics*, 26(19):2460–2461, 2010.

[6] Isak Gath and Amir B. Geva. Unsupervised optimal fuzzy clustering. *IEEE Transactions on pattern analysis and machine intelligence*, 11(7):773–780, 1989.

[7] Sarra Ben Hariz, Zied Elouedi, and Khaled Mellouli. Clustering approach using belief function theory. In *International Conference on Artificial Intelligence: Methodology, Systems, and Applications*, pages 162–171. Springer, 2006.

[8] Mylène Hugoni, Najwa Taib, Didier Debroas, Isabelle Domaizon, Isabelle Jouan Dufournel, Gisèle Bronner, Ian Salter, Hélène Agogué, Isabelle Mary, and Pierre E Galand. Structure of the rare archaeal biosphere and seasonal dynamics of active ecotypes in surface coastal waters. *Proceedings of the National Academy of Sciences*, 110(15):6004–6009, 2013.

[9] Weizhong Li, Limin Fu, Beifang Niu, Sitao Wu, and John Wooley. Ultrafast clustering algorithms for metagenomic sequence analysis. *Briefings in bioinformatics*, page bbs035, 2012.

[10] Weizhong Li and Adam Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, 2006.

[11] Frédéric Mahé, Torbjørn Rognes, Christopher Quince, Colomban de Vargas, and Micah Dunthorn. Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ*, 2:e593, 2014.

[12] Airel Pérez-Suárez, José F Martínez-Trinidad, Jesús A Carrasco-Ochoa, and José E Medina-Pagola. Oclustr: A new graph-based algorithm for overlapping clustering. *Neurocomputing*, 121:234–247, 2013.

[13] Torbjørn Rognes, Tomáš Flouri, Ben Nichols, Christopher Quince, and Frédéric Mahé. Vsearch: a versatile open source tool for metagenomics. *PeerJ*, 4:e2584, 2016.

[14] Simon Roux, Michaël Faubladier, Antoine Mahul, Nils Paulhe, Aurélien Bernard, Didier Debroas, and François Enault. Metavir: a web server dedicated to virome analysis. *Bioinformatics*, 27(21):3074–3075, 2011.

[15] Sarah L Westcott and Patrick D Schloss. De novo clustering methods outperform reference-based methods for assigning 16s rrna gene sequences to operational taxonomic units. *PeerJ*, 3:e1487, 2015.