Recursive Clustering Using Different Features Sets for Metagenomic Data

Isis Bonet[1*], Widerman Montoya[1], Andrea Mesa-Munera[1], Juan Fernando Alzate[2]

[1] EIA University, km 2 + 200 Vía al Aeropuerto José María Córdova, Envigado, Antioquia, Colombia.
[2] Centro Nacional de Secuenciación Genómica-CNSG, Facultad de Medicina, Universidad de Antioquia, Calle 67 Número 53-108, Medellín, Antioquia, Colombia.

* Corresponding author. Tel.: 57-43549090; email: ibonetc@gmail.com
Manuscript submitted October 17, 2017; accepted January 10, 2018.
doi: 10.17706/jcp.13.8.905-912

Abstract: Metagenomics binning process is a step prior to the taxonomic assignment of metagenomic reads of contigs, which helps to group genome sequences belonging to the same species. In this paper we propose a clustering method that is executed recursively to cluster contigs into groups of same taxa. In each step the method increases the taxonomic level, beginning with a domain and ending with a group that represents the species. The method uses a previous rule-based system to separate virus from the rest of the organism and feature selection algorithms to select different features in each step of the clustering. The clustering is based on *k*-means++ using Cosine and Jaccard distance, and feature selection on gain information. The proposed method outperforms classic *k*-means++, achieving 88.15% of purity in clusters.

Key words: Binning process, clustering, feature selection, metagenomics.

1. Introduction

Metagenomics is a breakthrough in sequencing the DNA, supported by modern next-generation sequencing technology. This research area is about the study of uncultured microorganisms obtained directly from environmental samples [1]. The low cost and the parallel throughput capacity of these technologies produce millions of sequenced DNA fragments (reads) [2]. An assembly process is executed to overlap reads and obtain longer DNA sequences called contigs. Although these sequences are longer than original reads, they are still too small to be compared with known organisms. Although a public big database of complete genome sequences is available, this represents only a small percentage of the biological diversity in our world. Thus, genome assembly and annotation is a significant challenge in metagenomics. The reconstruction of genomes in a specific environment from short DNA fragments can be compared to having a mixture of different puzzles, and then separate and assemble them. This task has an additional complexity, which consists in determining the types of organisms and the numbers of each type [3], [4].

Binning is the process of grouping reads or contigs and of associating them to a specific taxon. The algorithm that has been developed for binning metagenomic fragments can be divided into two categories: supervised and unsupervised methods. Supervised binning consists in the alignment of sequences, using databases of known organisms [5]. Examples of the most popular algorithms based on homology include BLAST [6], MEGA [7] and CARMA [8]. Other groups of supervised algorithms are based on composition and use a classification method such as k-Nearest Neighbor [9], Support Vector Machine [10] or Naïve Bayes

[11].

Unsupervised binning is based on composition. This means that composition features are required to represent the sequences. Some of the most common features that are used are GC-content and k-mers. TETRA is one unsupervised method, which uses 4-mers as features [12].

In this paper we propose a clustering method for binning contigs, which recursively builds groups, based on different features and representing a different taxonomic level in each clustering iteration.

2. Data and Methods

2.1. Data and Features

We use assembled genomic sequences at a contig level of a total of 16 organisms, including 5 viruses (HIV, Chikungunya, Ebola, Influenza and Dengue), 2 bacteria (Bacteroides dorei and Bifidobacterium longum) and 9 eukaryotes (Ascaris suum, Aspergillus fumigatus, Bos taurus, Candida parasilopsis, Glossina morsitans, Malus domestica, Manihot esculenta, Pantholops hodgsonii and Zea mays). This data was extracted from a publically accessible sequenced DNA, which is available on the FTP site of the Sanger institute (ftp://ftp.sanger.ac.uk).

This heterogeneous database consists of 872576 contigs in total, which vary in size between 50 and 30000 nucleotides bases.

For feature representation, we use a composition-based. The selected features are:

- Nucleotide frequencies: Number of occurrences of A, T, G and C in the sequence. They were normalized by the size of the sequence.
- GC content percentage: Percentage of G and C in the sequence.
- Codon frequencies or 3-mer: Number of each possible codon in the sequence. It was normalized by the total of codons (64 codons).
- *k*-mers (*k*=4), which is a 4-combination of the nucleotides. That means words of 4 letters [13]. The 4-mers are computed as the number of each tetranucleotide and normalized considering the total of tetranucleotides in the contig.

2.2. Methods

For unsupervised problems different clustering methods have been proposed. In this work we use *k*-means++ [14], which is a variant of *k*-means [15] that improves the selection of the centroids for each cluster. This algorithm finds a set of *k* centroids based on a weighted probability distribution, where a point *x* is chosen with a probability proportional to a distance function. This selection ensures that the centroids are distant from one another. After the centroids are chosen, the algorithm proceeds applying the standard *k*-means clustering.

In this research we use Euclidean and Cosine as distance functions.

2.3. Validation

There are different validation measures for clustering, but here, to assess the final quality of the clustering method, we use the purity of the clusters (1).

$$Purity(C_j) = \frac{\max(n_{ij})}{n_j} \tag{1}$$

where n_j is the number of organisms in cluster j (C_j) and n_{ij} is the number of organisms of class i in cluster j.

To measure the compactness of the clusters an internal evaluation, which is based on the intra-cluster distance is used.

For the implementation of the proposed clustering method, we used the Weka software in version 3.9

[16], which is a free machine learning package that provides the *k*-means++ algorithm. Furthermore, it has the advantage that it is easy to add a new clustering method.

3. Sequential Clustering with Different Features and Groups

The objective in metagenomics is to classify each contig into the species to which it belongs. From the perspective of biological sciences, different taxonomic classification can be established, starting with the domain (Bacteria, Eukarya), including virus, the highest taxonomic rank, and terminating with species, which is the basic unit of biological classification.

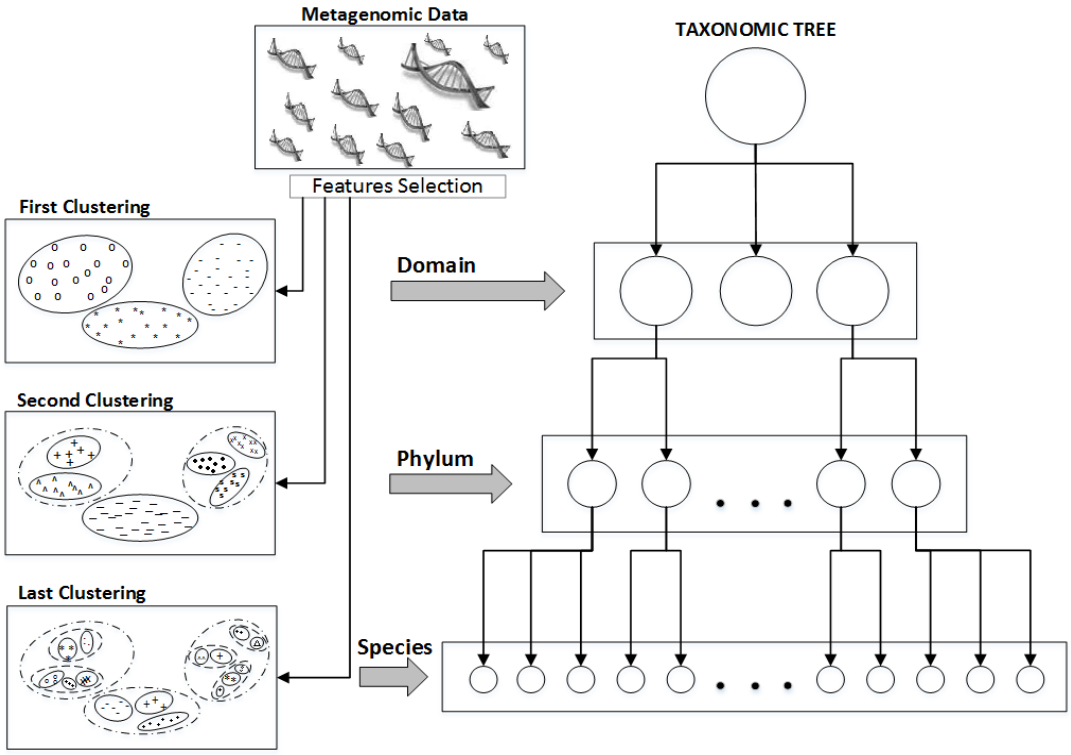


Fig. 1. Recursive clustering model with feature selection.

Accordingly, the proposed method is based on a real classification, increasing the taxonomic level. First, the metagenomic sequences are categorized into groups that represent the possible domains. Then, each separated group, that is a result of the previous segmentation process, is regrouped into the possible phylum. Finally, we regroup into species.

The assumption is that different features match with each taxonomic level. Fig. 1 represents the proposed method, which is based on the use of different features for each clustering iteration; to recursively regroup the organisms into smaller clusters that represent an operational taxonomic unit (OTU). Eventually, a tree will be built that organizes the different species in the taxonomic ranks. The first level of the tree represents the domain, the second the phylum and the last rank stands for the species. The sequenced fragments can be organized into groups of the same organism, in sequential steps increasing the specific level. This holds true even when, due to the given characteristics of the applied unsupervised method, the node of the tree does not specifies the specie.

Taking into account that viruses are a group relatively different from the other two domains, Bacteria and Eukarya, the codon usage of a set of these organisms was analyzed. Some peculiarities in that group can be stated. These can serve as a basis for the implementation of a rule-based system. A feature GC context

was chosen and a feature selection of codons or 3-mers completed.

After using the rule-based system to separate the viruses from the rest of organism, we consider the resulting group as the first cluster and perform a *k*-means clustering for the rest of the organisms with the objective to separate bacteria from eukaryotes.

A second step selects a new set of features that are present in each group to cluster again and to regroup the organisms into smaller clusters that represent the phylum-level taxa. The last step clusters the groups that were obtained in second step to obtain a possible representation of the species that form the sample.

The key of the method consist in the first separation of viruses from the rest of the organisms and the in applying recursively the clustering methods in order to obtain a more specific taxa level in each iteration, which provides in a first step the binning in metagenomics.

4. Results

In this paper a metagenomic database is used, which stores different species of different domains. The proposed recursive clustering method generates a tree, which represents each iteration of the clustering. That means each taxonomic level. The tree's leaves store the final result of the method in terms of the clusters that group the contigs into the OTUs assigned.

As already mentioned, the first step of the method consists of two parts. First, a rule-based system is built, using 5% of the organisms to perform a feature selection. As a result of computing the codon usage of these selected features it can be stated that some of the codons that most appear in viruses are: AAG, ACA, AGA, GGA, TCA, AGG, GAA, CAA, and GAT. In viruses, the observed percentage occurrence with which each of these codons is present in the contig is superior to 2.5%. Based on the analysis of the relation between these features, our proposal is to classify each contig as a virus, if it has a combination of at least three of the codons mentioned before.

In addition, a high value of GC content was observed in virus contigs, superior to 40%. Bearing in mind these two results, we built a rule system that connects with a simple rule both conditions. The obtained results by using this rule-based system show an error in viruses of 5% and in eukaryotes and bacteria of 15%. So far, the database is divided into two datasets, one stores most viruses and the other the eukaryotes and bacteria.

Subsequently, these two datasets is used as databases for two independent clustering processes that executes *k*-means++. First of all we used a feature selection method that is exclusively based on gain information with codons, and selected the rankest features. As a distance function the Cosine, Euclidean and Jaccard distance are used.

The results of separating contigs into the first taxonomic level are shown Table. 1. These results reveal that even when the second cluster contains 99% of bacteria, it holds 20% of eukaryotes, but eukaryotic contigs represent 99.7% of the total cases in the database.

Table 1. Results of the First *k*-means Clustering

Cluster_Domain	Purity	% of Eukaryotes	% of Bacteria
Cluster 1 (Eukarya)	99.99%	80%	1%
Cluster 2 (Bacteria)	1.22%	99%	20%

At this stage we have three subsets of data, where each one represents a domain: bacteria, eukarya and virus.

The idea for second step is to divide each of these subsets into smaller subsets associated with phylum

taxa and for the last clustering is to divide into the smallest clusters representing species.

The strategy here is to overestimate the number of clusters in order to obtain pure clusters. It is more important to obtain a cluster with a simple organism than with all contigs of the same species, even when the group is incomplete. Due to the fact that it is an unsupervised method, it is recommended that an expert defines the number of clusters. Here, twice as many species as total amount are selected, with a deviation of 2. This means, for example for a virus subset k= 10±2 species are selected. To decide best results, we compute the intra-cluster distance and select the most compact group of clusters.

During the course of the last clustering gain information is also used for feature selection, but now the method with 4-mers is applied. Finally, each cluster is assigned to the species which is most frequent in the cluster.

Fig. 2 illustrates the final clustering results. It shows the percent purity of clusters and the majority species for each cluster. The best results for virus were obtained using the Cosine distance and with k=12, for bacteria applying the Jaccard distance and with k=5, and for eukaryotes with k=18 and the Euclidean distance. In total 35 clusters were obtained, yielding 88.15% of purity. For virus the result is 93.5%, for bacteria 90.4% and for eukaryotes it is 83.95%. Among the species that could not be completely separated from the rest are Bifidobacterium longum, Ebola, Aspergillus_fumigatus and Candida parasilopsis.

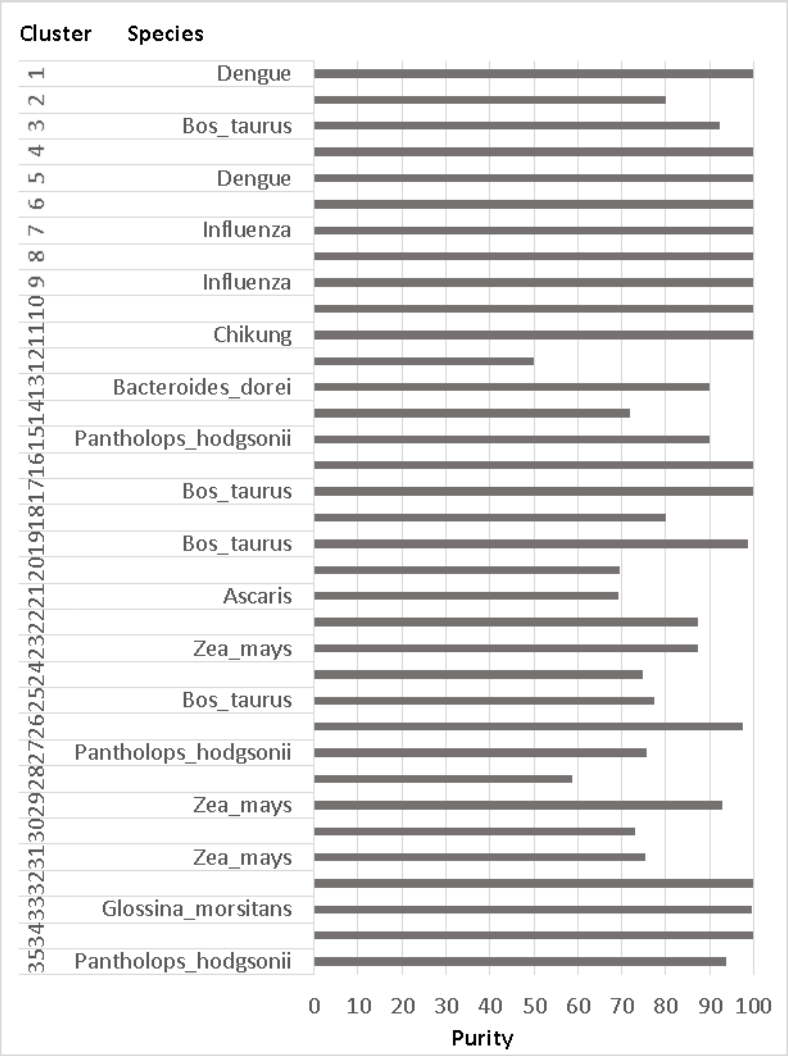


Fig. 2. Purity of clusters obtained with proposed recursive clustering method and the majority species for each cluster.

In essence, the application of a rule-based system for separating the viruses improves the results of clustering, as it is easier in a smaller subset of contigs to separate them from the rest. It is still a problem to separate Bacteria from eukaryotes. However, the last clustering was more effective with respect to the separation of species. Even when some clusters of the taxonomic tree do not belong to the right domain and phylum, the leaves, which represent the species, show a good result for the percentage of purity.

The last clustering results improved the results, which were obtained with *k*-means++, for all species in the same database. This demonstrates that it is easier to separate species from the same domain. Although the imbalance of the database also influences in the accomplishment of the objective to separate the species, as it can be noticed in the case of the bacteria. Commonly, the DNA of eukaryotes is larger and consists of more contigs. Furthermore, some bacteria can be much larger than others, as it happened to be in the present study.

5. Conclusions

This paper presents a method that applies recursive clustering, with a previous separation of virus genomes, by using a rule-based system. In each iteration the method generates clusters that are associated to different taxa levels. The difference in each clustering iteration refer to the features that were used in training them. The data are reconstructed using different features, which were selected previously by using gain information as the feature selection method.

It can be concluded that it is a favorable method that should be considered in the binning process as a previous step in the taxonomic assignment process. The proposed method was applied to a metagenome database composed of viruses, bacteria and eukaryotes. The result that was obtained by executing the proposed method outperforms the result that can be achieved with a simple clustering method. With regard to purity, it can be concluded that the proposed method provides pure clusters, reaching 88.15% of purity.

In future work, this recursive clustering can be used with other base clustering methods, such as SOM or Expectation Maximization, and with other distance functions.

References

[1] Chen, K., & Pachter, L. (2005). Bioinformatics for whole-genome shotgun sequencing of microbial communities. *PLoS Computational Biology*, *1(2),* 106-112.

[2] Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, *24(3),* 133-141.

[3] Pop, M., & Salzberg, S. L. (2008). Bioinformatics challenges of new sequencing technology. *Trends in Genetics*, *24(3),* 142-149.

[4] Eisen, J. A. (2007). Environmental shotgun sequencing: Its potential and challenges for studying the hidden world of microbes. *PLoS Biology*, *5(3),* 384-388.

[5] Thomas, T., Gilbert, J., & Meyer, F. (2012). Metagenomics — A guide from sampling to data analysis. *Microbial Informatics and Experimentation*, *2(3).*

[6] Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics, 10(1)*, 421-429.

[7] Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Research, 17(3),* 377-386.

[8] Gerlach, W., Jünemann, S., Tille, F., Goesmann, A., & Stoye, J. (2009). WebCARMA: A web application for the functional and taxonomic classification of unassembled metagenomic reads. *BMC Bioinformatics, 10(1),* 430.

[9] Diaz, N. N., Krause, L., Goesmann, A., Niehaus, K., & Nattkemper, T. W. (2009). TACOA — Taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. *BMC Bioinformatics, 10*, 56-71.

[10] McHardy, A. C., Martin, H. G., Tsirigos, A., Hugenholtz, P., & Rigoutsos, I. (2007). Accurate phylogenetic classification of variable-length DNA fragments. *Nature Methods, 4(1),* 63-72.

[11] Rosen, G., Garbarine, E., Caseiro, D., Polikar, R., & Sokhansanj, B. (2008). Metagenome fragment classification using N-mer frequency profiles. *Advances in Bioinformatics.*

[12] Teeling, H., Waldmann, J., Lombardot, T., Bauer, M., & Glockner, F. (2004). TETRA: A web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics, 5(1),* 163-169.

[13] Bonet, I., Montoya, W., Mesa-Múnera, A., & Alzate, J. (2014). Iterative clustering method for metagenomic sequences. In R. Prasath, P. O'Reilly, & T. Kathirvalavakumar (Eds.), *Mining Intelligence and Knowledge Exploration*. Springer International Publishing.

[14] Arthur, D., & Vassilvitskii, S. (2007). K-Means ++: The advantages of careful seeding. *Proceedings of the 8th Annual ACM-SIAM Symposium on Discrete Algorithms*, New Orleans.

[15] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability* (pp. 281-297).

[16] Witten, I., & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques* (2nd ed.). San Francisco: Morgan Kaufmann.

Isis Bonet was born in Cuba on July 5, 1978. She received the B.Sc degree (with honors) in computer science from the Universidad Central "Marta Abreu" de Las Villas (UCLV), Santa Clara, Cuba, in 2001, her M.Sc degree in computer science at UCLV in 2005 and her Ph.D in technical sciences (computer science) at UCLV in 2009.

She was member of the Bioinformatics and Artificial Intelligence Labs of the Center of Studies on Informatics, UCLV. She is currently with the EIA University, Colombia. Her research interests include artificial intelligence, business intelligence, data analytics, machine learning and bioinformatics.

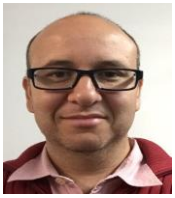
Bonet has authored/coauthored 45 papers in conference proceedings and scientific journals. She earned Cuban Academy of Sciences Award in 2011 and National Award of Commission of Scientifics Women of Cuba to the best scientific results in 2012.

Widerman Montoya was born in Colombia on January 31, 1992. He studied informatics engineering in the "EIA University" and got his B.Sc degree (with honors) 2014, awarded by his thesis in the study of clustering methods focus in metagenomic problems.

Montoya is working in a multinational company dedicated to the personal care in the business intelligence and data analytics area, and he is also working at EIA University.

Andrea Mesa-Munera was born in Medellín, Colombia on March 9, 1984. She studied systems engineering and computer science at the National University of Colombia — Medellín, graduated in June 2006, then continued with a master degree in engineering — systems engineering at the same University, where she was awarded with a scholarship. She obtained the title of magister in July 2009 and won an honorable mention for her master's thesis named "Method for handle of load balancing in computing distributed systems of high performance". She made an investigative internship in Mexico at the CINVESTAV of the IPN – Guadalajara.

After completing her master degree, she began to work as a teacher at the EIA University, Colombia in Febrary 2010. She is still working there.



Juan F. Alzate was born in Medellin, Colombia on August 5th, 1974. He studied microbiology at Universidad de Antioquia, Medellin, Colombia, graduated in 1996 (B.Sc), then continued with a master's degree in tropical parasitology in the same University. Ph.D awarded in 2006 at the Universidad de Alcala in Madrid, Spain, with a thesis dedicated to study programmed cell death in protozoan parasites.

When he returned to Colombia in 2006, he started to work as an assistant professor in the School of microbiology during two years. In 2008 when won fixed position in the same university and continued working as a full time professor at the Department of Microbiology and Parasitology, School of Medicine, Universidad de Antioquia.

Alzate has authored/coauthored more than 50 papers in conference proceedings and scientific journals. In 2009 he was awarded with a national grant to start the first reference center for genomics in Colombia, located at the Sede de Investigacion Universitaria – SIU, Universidad de Antioquia.