Iterative Clustering Method for Metagenomic Sequences

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

[1] Escuela de Ingeniería de Antioquia, Envigado, Antioquia, Colombia
{ibonetc,widerman.montoya,amesamu}@gmail.com
[2] Centro Nacional de Secuenciación Genómica-CNSG, Facultad de Medicina, Universidad de Antioquia
jfernando.alzate@udea.edu.co

Abstract. Metagenomics studies microbial DNA of environmental samples. The sequencing tools produce a set of genome fragments providing a challenge for metagenomics to associate them with the corresponding phylogenetic group. To solve this problem there are binning methods, which are classified into two sequencing categories: similarity and composition. This paper proposes an iterative clustering method, which aim at achieving a low sensitivity of clusters. The approach consists of iteratively run *k*-means reducing the training data in each step. Selection of data for next iteration depends on the result obtained in the previous, which is based on the compactness measure. The final performance clustering is evaluated according with the sensitivity of clusters. The results demonstrate that proposed model is better than the simple *k*-means for metagenome databases.

Keywords: Metagenomics, clustering, sequences binning, *k*-means.

1 Introduction

Metagenomics is a new science that combines different research field as genomics, bioinformatics and system biology. The objective of this field is to study genomes of many microbial organisms from a specific environment, which cannot be cultivated in laboratory. Understanding microbial communities' structure is a challenge in different areas such as biomedical, agriculture, environmental and life sciences [1].

The fast development of DNA sequencing techniques using different technologies generations, such as GS-FLX (454) /Roche, Solexa /Illumina, ABI SOLID /Applied Biosystems of second generation, and Helicos TSMS / Helicos BioSciences, Pacific BioSciences /Pacific BioSciences, of third generation; has led to new challenges in metagenomic studies [2]. Such studies are looking for identify the microorganisms in a sample to determine its metabolic functions [2]. Sequencing tools produce a puzzle of sequence fragments, which are known in this field with the name of reads (genome fragments). Following studies of the reads are performed, with the purpose of assembling them by a process of overlapping using large sequences named contigs [3].

* Corresponding author.

R. Prasath et al. (Eds.): MIKE 2014, LNAI 8891, pp. 145–154, 2014.

There is an even bigger problem resulting from the contigs that is related with the ongoing assembly process to obtain the complete genome, because of the analysis of an environmental sample contained diverse individuals.

Metagenomics also requires a binning process that allows contigs assignment of different species to their corresponding phylogenetic group. There are some methods of binning, which are classified into two sequencing categories: similarity and composition. Within this category it's found different software such as BLAST [4] and Phylopythia [5]. MEGAN [6] is one of the most widely used binning methods based on similarity sequence, which assigns reads to taxa, based on BLAST results. These kinds of methods are supported by a database of known species genome and they use similarity techniques as alignment to find similar sequences. For this reason, binning algorithms based on similarity are very time consuming. On the other hand, binning methods based on composition sequence made analyzes of genomes features, such as GC content, codon usage or oligonucleotide frequencies to describe the sequences and find clusters that represent the different taxonomic groups. Other features commonly used are called *k*-mers and represent the characteristic of oligonucleotide frequency of fragments sequence with size k, so that to compare them with a reference set of complete genomes [7]. The *k*-mer feature, with k=4, is widely used, also known as tetranucleotide frequencies. For example, TETRA method yields a statistical analysis using tetranucleotide patterns based on the characteristic of the GC content [8].

Binning algorithms based on similarity need this kind of features to compare the sequences between each other and grouped them based on their similarity into different clusters in order to seek the taxonomic groups in the data sample. Leading with this problem some author had been used different strategies; one of them is used clustering methods. In [9] a Self-Organizing Maps (SOM) method was used for efficiently cluster complex data using the oligonucleotide frequencies calculation. MetaCAA also is a clustering method based on tetranucleotides frequencies [3]. In [10] a comparison of some clustering methods is done.

Binning methods using a complete genomes knowledge-based classifier are referred to supervised learning methods, while methods that do not depend on training data are referred to unsupervised learning methods. Unsupervised learning methods are focused on major classes of collected data and do not perform well with data samples that don't have a significant population. On the other hand, supervised learning methods have a better performance in classifying the data of small populations [11].

In metagenomics, supervised learning methods are more precise, but they are time consuming because of the amount of different organisms present in the sample. Reducing organisms in the sample can improve their performance. That means, if binning method using knowledge-based classifier gets a set of subsequences of the same organism as input, the process to find the specify organism is easier and faster. A previous clustering process can be a way to provide different groups as inputs for supervised learning methods of binning. However, the aim is to find a clustering method which builds pure clusters. That is, members of each cluster belong to the same organism. This doesn't mean all subsequences of one organism are in the same cluster.

This paper is focused on an unsupervised method for assignment of genomic fragments into pure clusters based on composition sequence. Some of the widely-used

sequence-based measures, such as GC content nucleotides usage and *k*-mers frequencies, have been used to represent the genomic fragments. Further, for clustering fragments to cluster that represent the different genomes in the sample, a clustering iterative process based on *k*-means is proposed. The method has several iterations in the subset of data with more "error", that is the instances that belong to less compact clusters. For each iteration of the method, the improvement of the compactness of clusters is shown.

2 Methods and Data

2.1 Data

Assembled genomic sequences at contig level of different organisms including viruses, bacteria and eukaryotes were downloaded from the FTP site of the Sanger institute as is shown in table 1.

Selected viral sequences include Influenza and Dengue virus genomes. Sixty four dengue genomes ranging from 10,785 to 10,392 bp and an average GC content of 45.95%. Eight influenza genomic sequences that ranged between 2309 and 853 bp and an average GC content of 43.06%. No ambiguous "N" nucleotides were present in these contigs.

Bacterial sequences come from Bacteroides dorei and Bifidobacterium longum. For B. dorei, a total of 1948 contigs that summed 6,771,958 bases was analyzed. The contig N50 calculated value was 11,054 bases and only 8 "N" ambiguous bases were present. The largest contigs have 83484 bases. For B. longum, a total of 2,377,370 bases contained in 33 contigs that ranged between 580,034 and 540 bases were analyzed. The calculated contig N50 value was 154,900 and no ambiguous "N" bases were detected. The GC content was 42.3% for B. dorei and 59.93% for B. longum.

Table 1. Sequences and data source

Organism	Data source
Aspergillus fumigatus	ftp://ftp.sanger.ac.uk/pub/project/pathogens//A_fumigatus/AF.contigs.031704
Ascaris suum	ftp://ftp.sanger.ac.uk/pub/project/pathogens//Ascaris/suum/genome/assembly/contigs.fasta
Dengue	ftp://ftp.sanger.ac.uk/pub/project/pathogens//Dengue/Dengue.fasta
Glossina	ftp://ftp.sanger.ac.uk/pub/project/pathogens//Glossina/morsitans/Assemblies/tsetseGenome-v1.tar.gz
Bacteroides dorei	ftp://ftp.sanger.ac.uk/pub/project/pathogens//Bacteroides/dorei/D8/454LargeContigs.fna
Bifidobacterium longum	ftp://ftp.sanger.ac.uk/pub/project/pathogens//Bifidobacterium/longum/454LargeContigs.fna
Candida parasilopsis	ftp://ftp.sanger.ac.uk/pub/project/pathogens//Candida/parapsilosis/contigs/CPARA.contigs.fasta
Influenza	ftp://ftp.sanger.ac.uk/pub/project/pathogens//Influenza/Santiago_7981_06.fasta

The selected eukaryotes included 2 fungi, 1 nematode and 1 insect. The analyzed fungi were the mold Aspergillus fumigatus and the yeast candida parasilopsis. A total of 29,416,758 bases of A. fumigatus were analyzed. Theses sequences were contained in 344 contigs that ranged between 2,962,289 and 1,001 bases. The calculated N50 value was 1,120,772 bases and 3995 ambiguous "N" bases were detected. In the case of C. parasilopsis, a total of 13,265,923 bases contained in 1592 contigs were used. The calculated contig N50 value was 14,196 and the sizes ranged between 66,655 and 1,003 bases. 2919 ambiguous "N" bases were counted. The GC content was 49.55% for A. fumigatus and 38.86% for C. parasilopsis.

The analyzed nematode was Ascaris suum. A total of 527,713,826 bases contained in 138,557 contigs were analyzed. The contig N50 value was 8,524 and the count of ambiguous "N" was 7,668. The GC content was 37.89%.

The insect genomic sequences belong to Glossina morsitans fly. A total of 363,109,041 bases contained in 24,072 contigs that ranged between 538224 and 101 bases. The calculated contig N50 value was 49,769 and no ambiguous "N" bases were detected. The GC content was 34.12%.

2.2 Features

For the experiment some features were selected:

- GC: G + C content

$$GC = \frac{G + C}{A + T + G + C}$$

where A, T, G and C are the count of different nucleotides in the sequence.

- Nucleotide frequencies: Number of occurrences of A, T, G and C in the sequence. It was normalized by the size of the sequence.
- Codon frequencies: Number of each possible codon in the sequence. It was normalized by the total of codons (64 codons)
- *k*-mer (*k*=4): are represented for the 256 possible tetranucleotides. It was compute as the number of each tetranucleotide and normalized with the total of tetranucleotides in the sequence.

Features were used in all combinations, producing 15 databases.

2.3 *K*-means

K-means is one of the most popular clustering methods, despite the problem to estimate the parameter *k* (number of cluster). This algorithm finds a set of *k* centroids, and associates each instance in the data to the nearest centroid, based on a distance function [12]. Here we proposed a clustering method based on *k*-means.

Euclidean (Equation 1) and Cosine (Equation 2) distance were used to compare the sequences.

$$Euclidean(X,Y) = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2} \tag{1}$$

$$Cosine(X,Y) = 1 - \frac{\sum_{i=1}^{n}(x_i \times y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2} \times \sqrt{\sum_{i=1}^{n}(y_i)^2}} \tag{2}$$

Where X and Y are the instance to compare, with dimension N (features number), and x_i and y_i denote the i^{th} feature of X and Y respectively.

For the implementation of the clustering method, we used Weka [13], which is a free machine learning package that has implemented *k*-means. Furthermore, it has the advantage that it is easy to add a new clustering method.

3 Iterative Clustering Method

The process of clustering is based on the following steps:

Step 1: Select a tentative *k*, preferably a higher value than expected. Run k-means with the data.

Step 2: After getting the first set of clusters, they are evaluated based on measures of compactness and separation of clusters. Clusters with low separation between theirs centroids are merged into one. By other hand the compactness is used to divide the database, this means that clusters with low compactness are used to build the new database to repeat the clustering process returning to step 1.

Step 3: Once the process is stable, that means the compactness and separation are lower than a threshold, the last step is to minimize, if possible, the number of clusters. Clusters evaluation is repeated, for all clusters resulted of each iteration of *k*-means.

At the beginning of the process if necessary select the appropriated features and distance measure. For this problem we use the sensitivity of clusters to evaluate them.

In short, the general idea of this clustering method is seek clusters with a high sensitivity. In metagenomics the aim to assign the sequences to a phylum is associated with the sensitivity taking into account the phylum that best represents each cluster. That means the sensitivity is measured focus on the percentage that represents each organism in each cluster.

3.1 Performance Measures

There are some measures in the literature to assess performance of clusters. Here we use measures based on the pairwise difference of between and within-cluster distances. These measures are used to evaluate the cluster in each step of the proposed method and join similar clusters.

Furthermore, to assess the performance of clustering we focused on the final composition of the cluster, that is the number of different groups, and the purity of clusters, understanding by "pure cluster" a cluster with genomic fragments that belong to only one organism. Considering that clustering in metagenomics is a way to reduce the time-consumer of methods based on similarity of sequences it is more important getting clusters with a predominance of a phylogenetic group. Keeping this in mind, we use binomial estimator (equation 3) as a measure of sensitivity to evaluate the results. Although, this measure is used for binary problem, here we suppose the predominant sequences in each cluster as the positive, and the other as negative. The sensitivity is computed for each cluster meaning a range of pureness. The general sensitivity is computed by the average of all cluster sensitivity.

$$Sensitivity = \frac{number\ of\ positive}{total} \tag{3}$$

Each cluster is labeled with the organism which has the greatest number of sequences inside. The organism belonging to one cluster with different label is considered wrong. The sensitivity can be computed by cluster and average the results of other clusters.

4 Results and Discussion

In this paper a clustering method based on repeating a classical clustering algorithm (*k*-means algorithm) consecutively by a set of data composed of the “bad” clusters is proposed. A cluster is considered “bad” when its compactness is low.

A metagenome database built from 8 different organisms is used to evaluate the method.

Some different attributes are used to describe the sequences: GC content, nucleotides frequencies, codon frequencies and tetranucleotides.

Euclidean and Cosine distances were used for the *k*-means algorithms.

The first step was to select the best features to describe the data. This selection was focused on the result of a *k*-means with *k* between 5 and 15. The estimation was only based on the sensitivity of clusters. As explained before, our aim is to obtain pure clusters despite some organism can be divided in different clusters. Later, these clusters of genomic fragments can be classified using a supervised algorithm easier and faster. It is more important to have clusters with only one organism than to group genomic fragments of the same organisms together. For this reason the sensitivity, which here represents the percentage of the predominant organisms in each cluster, is a good measure for clustering.

The best result was obtained with *k*=15, tetranucleotides as features and Cosine distance. Figure 1 shows this result. The left part of the figure represents the number of clusters, the organisms assigned and the number of fragments associated with each organism. It can be seen most of clusters have a percentage relative to the predominant organism superior of 90%. The sensitivity was 92.85%, nevertheless the organism are very scattered.

Once selected the representation of data and the parameters of the *k*-means we test the proposed method. Starting from the result obtained before we go to the step 2 in order to evaluate the different clusters to merge closer clusters and separate clusters less compact for the next step. The process was repeated five times until to achieve the stability of the model.

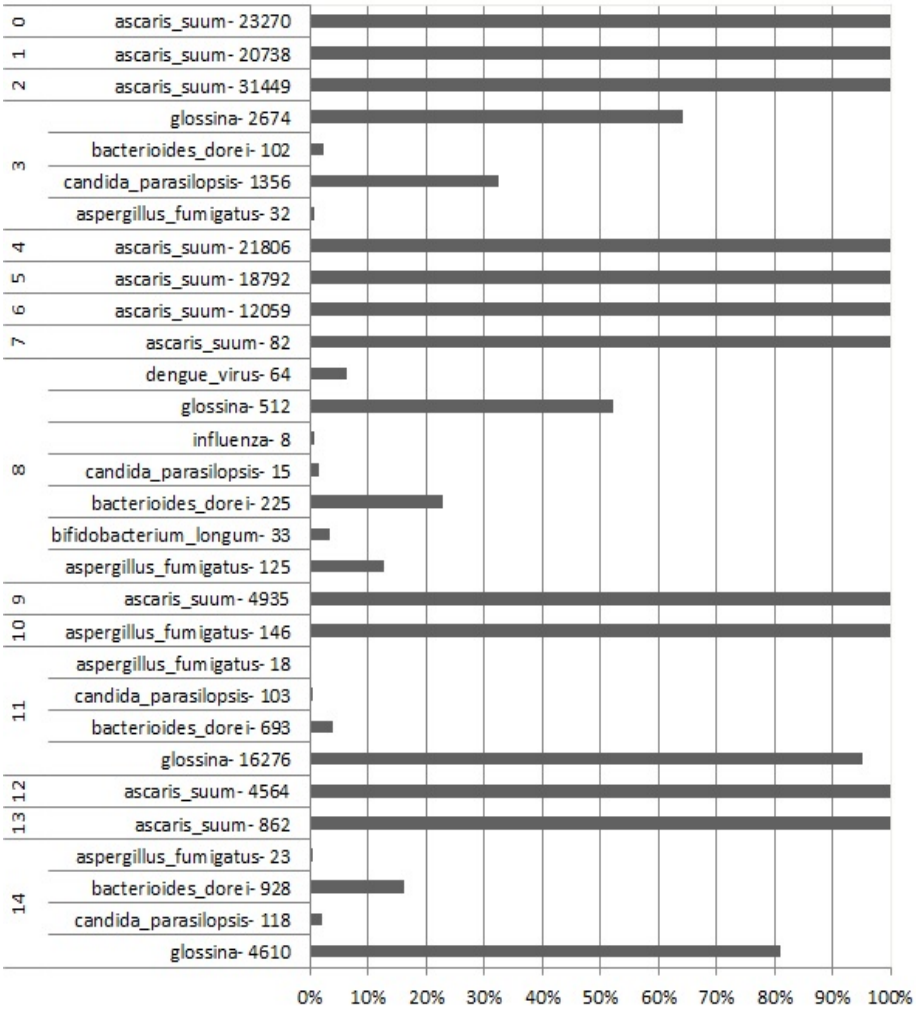

Fig. 1. Results of classical *k*-means with tetranucleotides frequencies as features

Final result is better compared with the first one (simple *k*-means). Figure 2 shows the results of the last step of the model yielding a 99.1% of sensitivity of the clusters, which results are in the range of 87.14 and 100%. The error of misassigned sequences is 5.516%.

With a more deep analysis of the results, we can see that *ascaris suum*, *bacterioides_dorei*, *bifidobacterium_longum* are completely grouped in clusters 0, 3 and 10 respectively. *Aspergillus_fumigatus* is grouped into 4 different clusters, but has two clusters complete for it. *Dengue* is divided into two groups. *Glossina* and *Candida* are more partitioned, although, *Glossina* leads all the clusters to which it belongs. The worst result is for *Influenza* because it is never recognized and separated of the rest organism, though its fragments are grouped together in the same cluster.

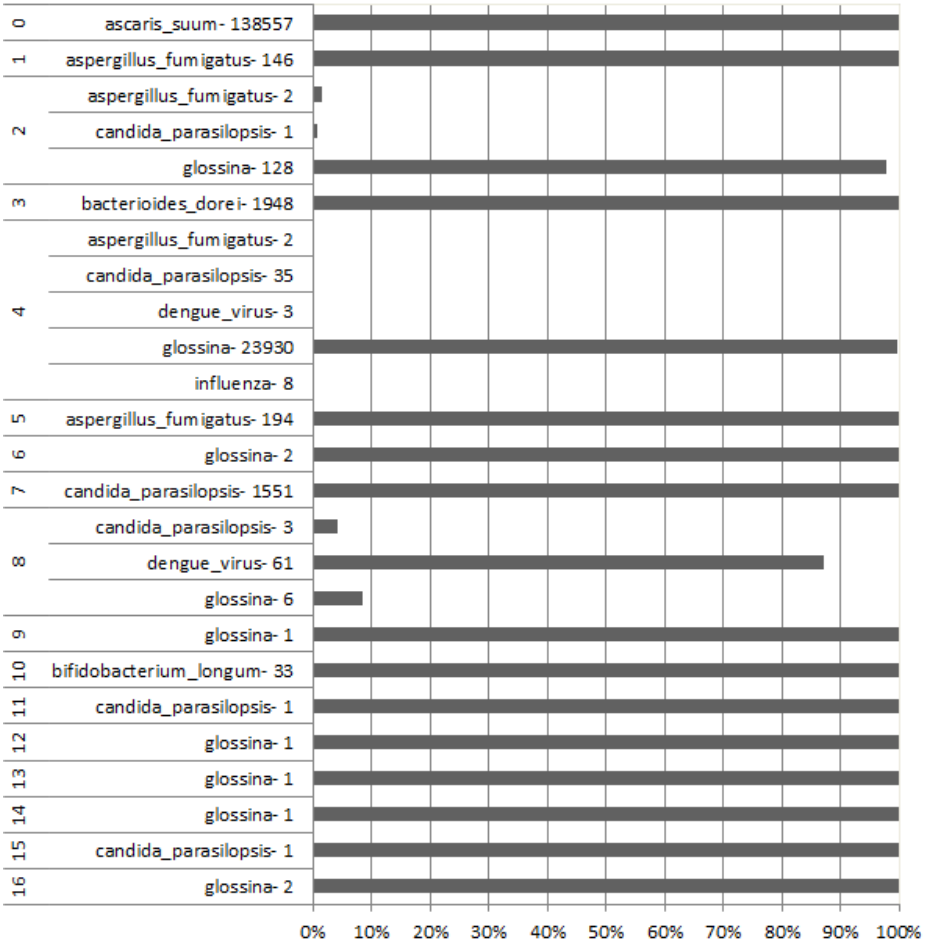

Fig. 2. Results of the proposed clustering method

In short, the results presented by the iterative use of *k*-means are superior of the only one running of *k*-means. By the application of metagenomics means an advantage this kind of group, although the patterns are divided in different group. Taking into account this, the error of the model based on the count of sequence misassigned is 0.045.

This paper is not intended to show the best clustering method for metagenomics, but rather to show a promising method to bear in mind for this area.

This iterative algorithm can be used with other base clustering method such as SOM or Expectation Maximization. In future work we expect compare the proposed method with other base methods and other metagenome databases.

5 Conclusions

In this paper we present an approach based on the iterative application of *k*-means to pattern that belongs to "bad" cluster. The classification of cluster is focused on validation measures of the compactness and separation cluster. The proposed method is applied to a metagenome dataset composed of 8 different organisms. The result achieved by the proposed method, in line with the objective of obtaining clusters with high sensitivity, outperforms result obtained with a simple *k*-means. Taking into account the error, the proposed method improves the purity of clusters by 5.471%. The results presented here do not mean that the method described here is better than other clustering methods for any metagenomic problem, but it is a promising method to bear in mind.

Other clustering methods can be used as the base for the proposed algorithm. This proposed method can also be applied to other metagenome databases.

References

1. Council, N.R.: The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet. The National Academies Press (2007)
2. Wu, Y.-W., Ye, Y.: A Novel Abundance-Based Algorithm for Binning Metagenomic Sequences Using *l*-Tuples. In: Berger, B. (ed.) RECOMB 2010. LNCS, vol. 6044, pp. 535–549. Springer, Heidelberg (2010)
3. Reddy, R.M., Mohammed, M.H., Mande, S.S.: MetaCAA: A clustering-aided methodology for efficient assembly of metagenomic datasets. Genomics 103, 161–168 (2014)
4. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.: BLAST+: architecture and applications. BMC Bioinformatics 10, 421 (2009)
5. McHardy, A.C., Martin, H.G., Tsirigos, A., Hugenholtz, P., Rigoutsos, I.: Accurate phylogenetic classification of variable-length DNA fragments. Nat. Meth. 4, 63–72 (2007)
6. Huson, D.H., Auch, A.F., Qi, J., Schuster, S.C.: MEGAN analysis of metagenomic data. Genome Research 17, 377–386 (2007)
7. Chan, C.-K., Hsu, A., Halgamuge, S., Tang, S.-L.: Binning sequences using very sparse labels within a metagenome. BMC Bioinformatics 9, 215 (2008)
8. Teeling, H., Waldmann, J., Lombardot, T., Bauer, M., Glockner, F.: TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. BMC Bioinformatics 5, 163 (2004)
9. Abe, T., Kanaya, S., Kinouchi, M., Ichiba, Y., Kozuki, T., Ikemura, T.: Informatics for Unveiling Hidden Genome Signatures. Genome Research 13, 693–702 (2003)
10. Li, W., Fu, L., Niu, B., Wu, S., Wooley, J.: Ultrafast clustering algorithms for metagenomic sequence analysis. Briefings in Bioinformatics 13, 656–668 (2012)

11. Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K., Hugenholtz, P.: A Bioinformatician's Guide to Metagenomics. Microbiology and Molecular Biology Reviews 72, 557–578 (2008)
12. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Statistics, vol. 1, pp. 281–297. University of California Press, Berkeley (1967)
13. Witten, I., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, San Francisco (2005)