ALIGNMENT-FREE METHODS FOR CLASSIFICATION OF METAGENOMIC DATA

Tereza Vaněčková

Master Degree Programme (2), FEEC BUT

E-mail: xvanec02@stud.feec.vutbr.cz

Supervised by: Helena Škutková

E-mail: skutkova@feec.vutbr.cz

Abstract: Metagenomics studies microbial communities by analyzing their genomic content directly sequenced from the environment. In this contribution, alignment-free methods based on word frequency will be introduced. It has been proven, that these methods are effective in processing of short metagenomic sequence reads produced by Next-Generation Sequencing technologies. To evaluate the potential of word frequency based methods, the *k*-mer analysis was applied on simulated dataset of metagenomic sequence reads with length of 600 nucleotides. Then the data were enrolled for a hierarchical cluster analysis. Results have shown that the proposed method is able to cluster genome fragments of the same taxa.

Keywords: metagenomics, alignment-free, nucleotide word frequency, hierarchical clustering

1. INTRODUCTION

With rapid development of sequencing technologies and applications, metagenomics is becoming an important approach for studying microbial communities in different enviroments and the human body. Metagenomic datasets consist of many raw reads that represent various parts of many individual genomes. [1]

The data volume generated by Next-Generation Sequencing (NGS) technologies is growing. Therefore, handling and processing such large datasets is becoming one of the major challenges in most metagenome research projects. [2] Alignment-based methods encounter difficulties in dealing with large datasets. To overcome this limitation many alignment-free methods have been proposed. Among them the method based on nucleotide word frequencies (*k*-mer analysis) may be the most developed alignment-free method. [3]

In this study the aim is to evaluate the ability of *k*-mer analysis approach to cluster reads of shotgun metagenomic data.

2. METHODOLOGY

Sequence signatures are normalized frequencies of *k*-mers in a sequence. Previous studies have shown that *k*-mer frequencies are similar across different regions of the same genome, but differ between genomes of different organisms. Therefore, they carry phylogenetic information. [3, 4]

2.1. NUCLEOTIDE WORD FREQUENCY

K-mer is a word that consists of *k* symbols from a finite alphabet of nucleotides A = {A, C, G, T}. The set W_k consists of all possible *k*-mers, also refers to variation with repetition (1). This set has *n* elements and the words are sorted alphabetically. [3]

$$W_k = \{w_{k,1}, w_{k,2}, \dots, w_{k,n}\} \qquad (1)$$

$$n = 4^k$$

Computationally, the counting of *k*-mers in a sequence of length *m* is usually performed by taking a sliding window of length *k*, that moves through the sequence from position 1 to *m* - *k* + 1. The overlapping of *k*-mers is allowed in this method. A sequence can be represented by *n*-dimensional vector c_k consisting of *k*-mers counts (2) [3]:

$$c_k = \left(c(w_{k,1}), c(w_{k,2}), \dots, c(w_{k,n})\right), \tag{2}$$

where *n* is the total number of possible *k*-mers. Vector of *k*-mer frequencies is counted as a relative number of each *k*-mer according to (3) [3]:

$$f_k = \left(f(w_{k,1}), f(w_{k,2}), \dots, f(w_{k,n})\right) \tag{3}$$
$$= \left(\frac{c(w_{k,1})}{m-k+1}, \frac{c(w_{k,2})}{m-k+1}, \dots, \frac{c(w_{k,n})}{m-k+1}\right).$$

2.2. HIERARCHICAL CLUSTERING

Hierarchical cluster analysis is widely used method for classification of features based on *k*-mer frequencies. [5, 6] It is a task of grouping a set of objects in such a way that objects in the same cluster are more similar to each other than to those in the other groups. Mutual similarity of sequences can be shown by creating a cluster tree or dendrogram. Ideally, branches of the dendrogram should contain a cluster of sequences of the same taxon. The hierarchical cluster analysis can be performed on a dataset by using following procedure [7]. First step is to find similarities or dissimilarities between every pair of objects (vectors of *k*-mer frequencies). Various distance metrics can be used with one of the most widespread Euclidean. Next, the objects are grouped into a binary hierarchical cluster tree with use of a linkage function (eg. nearest neighbor or furthest neighbor). The last step is to determine the number of clusters.

3. RESULTS

Analysis, as shown below in **Figure 1**, was performed on the simulated metagenomic dataset. This dataset consists of 60 genomic fragments (with length of 600 nucleotides) originating from 3 bacterial genomes of different bacterial phyla. For each of the fragments, the vector of *k*-mer frequencies (where *k*=7) was calculated. Based on the calculation of cophenetic correlation coefficient, the method of Spearman distance and furthest neighbor linkage function was determined as the most powerful method for clustering.

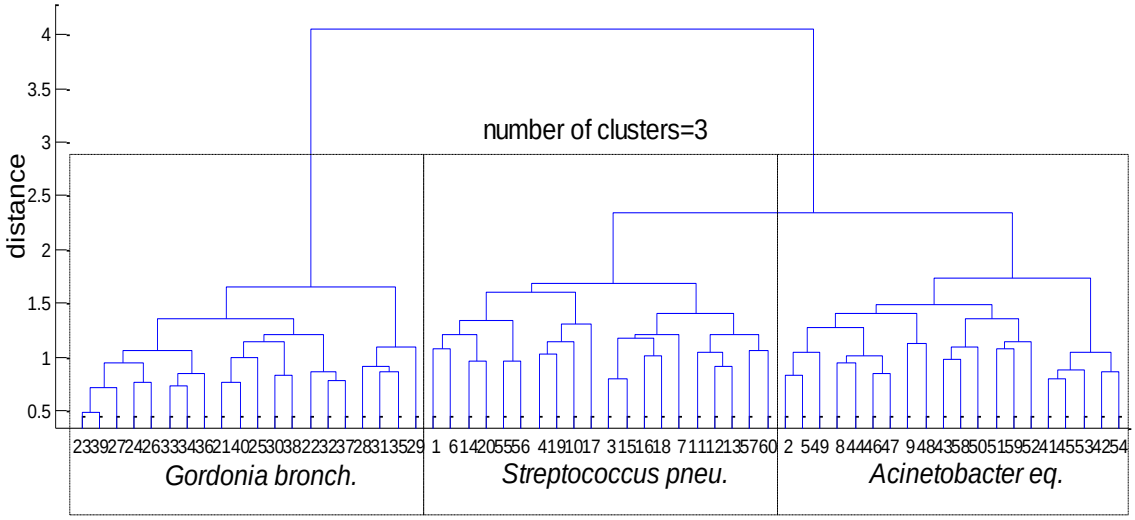

Figure 1: Analysis of 60 genomic fragments of 3 bacterial genomes. Rectangles represent clusters.

4. DISCUSSION AND CONCLUSION

To evaluate the potential of word frequency based methods for classification of metagenomic data, the *k*-mer analysis was applied on 60 simulated metagenomics sequence reads with length of 600 nucleotides. Then the data were enrolled for hierarchical cluster analysis.

As shown in the example (Figure 1), this method is able to separate sequences of taxonomically distant species, therefore to create clusters of genomic fragments originating from the same taxon. In this case, bacterial species of different bacterial phyla were used. The impact of *k*-mer length on method performance can be seen from the **Chyba! Nenalezen zdroj odkazů.** below. The number of misclassified fragments in analyzed dataset was 13 (78,3% success rate) when *k*=2. In contrast, when *k*=7, there was 8 misclassifications (86,6% success rate). Although there were found slightly better results with increasing length of *k*-mer, there is no point in counting *k*-mers longer than 7 in reads of length approximately 600 *bp* as no significant improvement in performance of the algorithm was found. The algorithm gives similar results to that publicated by Yang *et al*. [6], where rank orders of values of *k*-mer counts in a sequence are considered.

Length of *k*-mer	2	3	4	5	6	7
Success rate - this study [%]	78,3	76,6	76,6	80	78,3	**86,6**
Success rate - Yang *et al*. [%]	78,3	70	76,6	**86,6**	76,6	71,6

Table 1: Impact of *k*-mer length on success rate of the algorithms

The method proposed above has several unique features. First, this method is unsupervised, therefore it does not require any prior knowledge for the clustering or binning of the data. Next, no reference database is needed, as assessing similarity to existing genome sequence databases is very often limited by their incompleteness. Moreover, this method does not require sequence alignment, hence it is less computationally demanding.

However, further research regarding extracting sequence features based on *k*-mer frequencies and classification methods is needed. Limitation of this method may be lower accuracy in finding clusters in data that contain sequences of lower taxonomic rank, e.g. genus or family. One of the reasons can be high similarity of the *k*-mer frequency vectors between these sequences.

REFERENCES

[1] BASHIR, Y., S. PRADEEP SINGH and B. KUMAR KONWAR. Metagenomics: An Application Based Perspective. *Chinese Journal of Biology*. Hindawi Publishing Corporation, 2014. DOI: 10.1155/2014/146030.

[2] HODKINSON, B. P. and E. A. GRICE. Next-Generation Sequencing: A Review of Technologies and Tools for Wound Microbiome Research. *Advances in Wound Care*. 2015, 4(1), 50-58. DOI: 10.1089/wound.2014.0542.

[3] VINGA, S. and J. ALMEIDA. Alignment-free sequence comparison-a review. *Bioinformatics.* Oxford, UK. 2003, 19(4). ISSN 13674803.

[4] TEELING, H. *et.al*. Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environmental Microbiology*. Oxford, UK: Blackwell Science Ltd, 2004, 6(9), 938-947. ISSN 14622912.

[5] JIANG, B. *et al*. Comparison of metagenomic samples using sequence signatures. *BMC Genomics*. London: BioMed Central, 2012, (13). DOI: 10.1186/1471-2164-13-730.

[6] YANG, X. and T. WANG. A novel statistical measure for sequence comparison on the basis of k-word counts. *Journal of Theoretical Biology*. 2013, (318), 91-100. DOI: 10.1016/j.jtbi.2012.10.035. ISSN 00225193.

[7] GENTLE, J. E., L. KAUFMAN and P. J. ROUSSEU. Finding Groups in Data: An Introduction to Cluster Analysis. *Biometrics*. 1991, 47(2), ISSN 0006341x.