SIGNAL BASED FEATURE SELECTION FOR FAST CLASSIFICATION OF SEQUENCES IN METAGENOMICS

Karel Sedlář

Doctoral Degree Programme (3), FEEC BUT

E-mail: sedlar@feec.vutbr.cz

Supervised by: Ivo Provazník

E-mail: provaznik@feec.vutbr.cz

Abstract: The rapid development in DNA sequencing techniques brings completely new possibilities into metagenomics research. No longer is the whole metagenome sequencing an issue. On the other hand, this progress lies new demands on bioinformatics tools indented to process this kind of data. Unlike the amplicon based sequencing where every sequence represents a particular gene, the whole metagenome sequencing produce sequences that are random pieces of genomes in the metagenome. Therefore, the reference database for identification of these sequences cannot be used. Here, we present fast feature selection based on genomic signal processing for alignment-free classification of sequences in the metagenome.

Keywords: metagenomics, WGS, genomic signal

1. INTRODUCTION

According to estimates, the number of cells belonging to microorganisms located in a human body exceeds the number of human cells in a ratio of 10 to 1. Although most of these organisms do not cause any harm to the body, not inconsiderable part of them is associated with the occurrence of dangerous diseases and antibiotic resistance. Accurate and rapid identification of these organisms is therefore fundamental in the research of diseases for fast deployment of appropriate treatment and therefore in saving of human lives [1]. In the research of farm animals, it can help prevent significant economic losses through preventing diseases of entire herds.

The rapid development of DNA sequencing techniques is dramatically changing the way in which metagenomic research is conducted, i.e. research of the genomes of different organisms found in the common environment. Such studies are nowadays closely associated with the use of advanced bioinformatics techniques for processing biological sequences [2]. Compared to two metagenomic studies published in 2006, thousands of studies are produced in these days. Thus, continuous development of new and faster bioinformatics tools that will be able to handle this amount of data is necessary [3].Older but still used metagenomic approach stands on targeted amplicon sequencing of one selected representative gene [4]. Sequence identification is then performed by comparing the sequences against the reference database and data can be visualized using the dimensionality reduction techniques, usually analysis of principal co-ordinates (PCoA) using UniFrac metric [5,6]. The current trend, however, lies in shotgun sequencing of the entire metagenome. Unfortunately this approach brings several issues. Firstly, using this approach, one is not able to identify the sequences directly, because they represent different parts of the genome, for which there is still no complete reference database. Secondly, the number of sequences prevents sufficient comparison against a reference database even with the heuristic algorithm BLAST [7,8].

Current bioinformatics methods for metagenomic data processing are based on fast classification of sequences into clusters that represent the different types of organisms and whose subsequent identification is no longer an issue. However, there are several different approaches. The simplest of them use locally sensitive hash function and compare the two sequences locally by particular words, so called *k*-mers [9,10]. Advanced techniques are based on a selection of different feature vectors. In such cases the classification of metagenomic data is normally carried out using clustering techniques [11]. It is assumed that individual clusters will contain the feature vectors derived from the same organism and the overall architecture of the clusters will give us information about taxonomic diversity of studied species. Unfortunately, current techniques for creating such a specific feature vector are based only on the character processing techniques and are therefore very limited. On the other hand, a large selection of techniques to generate a sufficiently specific vector can be provided by a progressively evolving discipline of bioinformatics called genomic signal processing [12]. As it has been already deduced, phase signals are species specific and can be used for example for a sequence alignment [13] or comparison of species even after a massive downsampling [14]. Here, we present reconstruction of species specific feature vector based on combination of slopes of several different phase signal representations of genomic data.

2. MATERIALS AND METHODS

2.1. TEST DATASET

In order to test the proposed method, we created 3 test dataset containing 2,500 sequences each, derived from 5 different organisms representing 4 different bacterial species. The whole genome sequences were obtained from Genbank database at NCBI (http://www.ncbi.nlm.nih.gov/genbank/), see summary in Table 1. Sequencing reads were simulated as random fragments derived from complete genome sequences and 500 reads were generated from every genome. A half of the sequences were further modified as reverse complementary to better represent real sequencing data. Moreover, 3 dataset according to length of reads (500 nt, 1,000 nt and 5,000 nt) were prepared to examine the influence of read length on the succession of classification and to represent data from next-generation sequencing (NGS) as well as third-generation (TGS) sequencing platforms.

Table 1: Summary of organisms used for test dataset

species	organism	accession no.
E. coli	*Escherichia coli* UTI89	NC_007946.1
E. coli	*Escherichia coli* str. 'clone D i14'	NC_017652.1
C. C. ruddii	*Candidatus Carsonella ruddii* PV	NC_008512.1
G. obscurus	*Geodermatophilus obscurus* DSM 43160	NC_013757.1
R. prowazekii	*Rickettsia prowazekii* str. Dachau	NC_017051.1

2.2. SIGNAL REPRESENTATION

Although we have already proved the cumulated phase signal representing purine-pyrimidines (R-Y) and strong-week (S-W) ratio to be species specific, this feature was examined only on a large (whole genome) scale [15]. To examine the features on a smaller (reads) scale, we decided to use all 3 possibilities in which nucleotides {A, C, G, T} are assigned phases {π/4, -3π/4, 3π/4, -π/4} to present R-Y and S-W, {3π/4, -3π/4, π/4, -π/4} to present R-Y and amino-keto (M-K), {3π/4, -3π/4, -π/4, π/4} to present S-W and M-K. Moreover, we examined both, cumulated phase signals representing cumulative sum of nucleotides along sequences as well as unwrapped phase representing transition and transversions along sequences [16].

For every read, all 6 signals (3 cumulated phase and 3 unwrapped phase signals) were reconstructed and their slopes were computed. A feature vectors representing reads were constructed using those 6 values representing the slopes.

2.3. CLUSTERING AND STATISTICS

Feature vectors were clustered using Ward's hierarchical clustering technique with utilization of Euclidean metric. The resulting clusters were compared to real taxonomy and statistics for each organism in form of sensitivity (*sensitivity=TP/(TP+FN)*), specificity (*specificity=TN/(TN+FP)*), precision (*precision = TP/(TP+FP)*) and accuracy (*accuracy = (TP+TN)/(TP+FN+FP+FN)*) were computed.

3. RESULTS AND DISCUSSION

Firstly, we reduced the feature vectors by omitting the signals of unwrapped phases representing S-W and M-K. This kind of signals is identical to unwrapped phase representing R-Y and M-K from its definition and therefore it does not bring any additional information.

Secondly, by computing the above mentioned statistics for particular signal representations, the slope of cumulated phase signals representing R-Y and S-W were found to be unable to distinguish different species. Therefore, we omitted also this feature from the vector. The example of cumulated phase signals with (S-W and M-K) and without (R-Y and S-W) discriminative information is shown in Figure 1.

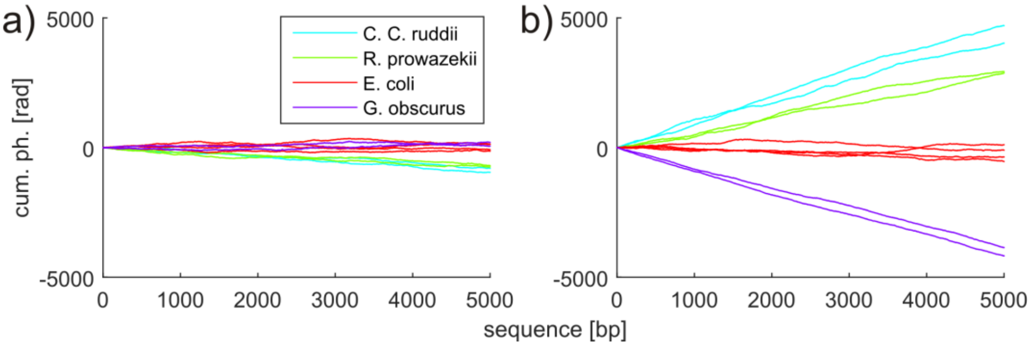

Figure 1: Cumulated phase signals for randomly selected sequences from the test dataset presenting a) S-W and M-K information b) R-Y and S-W information

The resulting vectors containing 4 slopes were further analyzed by hierarchical clustering and the resulting clusters were used for statistics evaluation. The tree reconstructed from feature vectors derived from 5,000 nt long sequences is presented in Figure 2. There are 4 evident clusters representing 4 species.

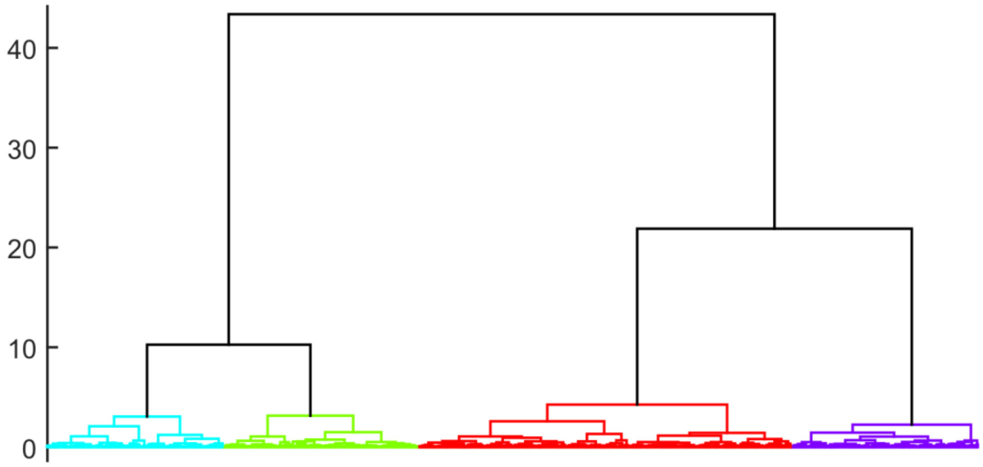

Figure 2: The tree reconstructed from the feature vectors using Ward method and Euclidean metric

The cluster containing *E. coli* vectors is larger because two organisms for this species were included in the test dataset. Only a small number of sequences were misclassified as shown below. However, the longer the sequences are, the higher amount of information they contain. Although 5,000 nt long reads are typical for TGS, we tried to test if the method works for NGS data too. Therefore, we repeated the analysis with shorter reads, typical for NGS. The results of classification are summarized in Table 2 and Table 3.

Table 2: The results of classification (sensitivity and specificity)

Organism	500 nt		1000 nt		5000 nt	
	Sensitivity	**Specificity**	**Sensitivity**	**Specificity**	**Sensitivity**	**Specificity**
E. coli	98,28	94,64	99,78	93,63	99,90	99,60
C. C.ruddii	75,00	98,14	97,45	97,98	100,00	99,21
G. obscurus	99,80	99,60	99,40	99,95	100,00	100,00
R. prowazekii	74,29	92,08	77,67	99,31	95,78	99,95
Average	**86,84**	**96,12**	**93,58**	**97,72**	**98,92**	**99,69**

Table 3: The results of classification (precision and accuracy)

Organism	500 nt		1000 nt		5000 nt	
	Precision	**Accuracy**	**Precision**	**Accuracy**	**Precision**	**Accuracy**
E. coli	91,60	96,00	89,80	95,84	99,40	99,72
C. C.ruddii	93,00	92,40	91,80	97,88	96,80	99,36
G. obscurus	98,40	99,64	99,80	99,84	100,00	100,00
R. prowazekii	67,60	88,84	97,40	93,88	99,80	99,08
Average	**87,65**	**94,22**	**94,70**	**96,86**	**99,00**	**99,54**

The results show that the method is very successful for TGS data classification, however, it is also applicable on NGS data with satisfactory results. The main advantage of the method over the current techniques is its low computational complexity. While the most of current techniques suffer from quadratic complexity and the fastest of them are no better than $O(n \log n)$ [17], our method is computationally very efficient, based only on basic mathematical operations. In combination with recently published algorithm *TwisterTries* for hierarchical clustering [18], the whole pipeline would be only $O(n)$.

4. CONCLUSION

A new method for classification of metagenomic data without prior alignment to a reference database is introduced in this paper. Unlike the current techniques, our approach relies fully on genomic signal processing making it computationally very efficient. The sequences are first transformed into 4 different phase signal representations and then the feature vector using the slopes of the signals is reconstructed. The only computationally demanding part remains in final vector clustering. However, this part can be updated with recently published algorithm for hierarchical clustering with only linear complexity. Such a method fully meets the current needs when millions of sequences are produced during every sequencing run. We demonstrated our method to work not only on TGS data, but also on currently most widely used NGS.

ACKNOWLEDGEMENT

Supported by BUT intern grant agency, interfaculty junior project FEKT/FIT-J-16-3335 Advanced methods for fast identification and visualization of metagenomic data.

REFERENCES

[1] NIH Human Microbiome Project [online]. 2016 [cit. 2016-03-10]. Dostupné z: http://www.hmpdacc.org/

[2] SIMON, C. a R. DANIEL. Metagenomic Analyses: Past and Future Trends. Applied and Environmental Microbiology. 2011, 77(4), 1153-1161.

[3] REDDY, T. B. K., et al. The Genomes OnLine Database (GOLD) v.5: a metadata management system based on a four level (meta)genome project classification. Nucleic Acids Research. 2015, 43(D1), D1099-D1106.

[4] KLINDWORTH, A., et al. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Research. 2012, 41(1), e1-e1.

[5] LOZUPONE, C. a R. KNIGHT. UniFrac: a New Phylogenetic Method for Comparing Microbial Communities. Applied and Environmental Microbiology. 2005, 71(12), 8228-8235.

[6] HAMADY, M., C. LOZUPONE a R. KNIGHT. Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. The ISME Journal. 2009, 4(1), 17-27

[7] SCHOLZ, M. B., C. LO a P CHAIN. Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. Current Opinion in Biotechnology. 2012, 23(1), 9-15.

[8] ALTSCHUL, S. F., W. GISH, W. MILLER, E. W. MYERS a D. J. LIPMAN. Basic local alignment search tool. Journal of Molecular Biology. 1990, 215(3), 403-410.

[9] EDGAR, R. C. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010, 26(19), 2460-2461.

[10] RASHEED, Z. a H. RANGWALA. A Map-Reduce Framework for Clustering Metagenomes. In: 2013 IEEE International Symposium on Parallel. IEEE, 2013, s. 549-558.

[11] MANDE, S. S., et al. Classification of metagenomic sequences: methods and challenges. Briefings in Bioinformatics. 2012, 13(6), 669-681

[12] ANASTASSIOU, D. Genomic signal processing. IEEE Signal Processing Magazine. 2001, 18(4), 8-20.

[13] SKUTKOVA, H., M. VITEK, K. SEDLAR a I. PROVAZNIK. Progressive alignment of genomic signals by multiple dynamic time warping. *Journal of Theoretical Biology*. 2015, 385, 20-30.

[14] SEDLAR, K., H. SKUTKOVA, M. VITEK a I. PROVAZNIK. Set of rules for genomic signal downsampling. *Computers in Biology and Medicine*. 2016, 69, 308-314.

[15] SEDLAR, K., H. SKUTKOVA, M. VITEK a I. PROVAZNIK. Prokaryotic DNA Signal Downsampling for Fast Whole Genome Comparison. *Advances in Intelligent and Soft Computing*, 2014, 373, 273-291.

[16] CRISTEA, Paul Dan. Large scale features in DNA genomic signals. *Signal Processing*. 2003, 83(4), 871-888.

[17] LACZNY, C., et al. Alignment-free Visualization of Metagenomic Data by Nonlinear Dimension Reduction. *Scientific Reports*. 2014, **4**(1), e4516.

[18] COCHEZ, M. a H. MOU. Twister Tries. In: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data - SIGMOD '15. New York, USA: ACM Press, 2015, 505-517.