Metagenome Fragment Binning Based on Characterization Vectors

Wisnu Ananta Kusuma*
Department of Computer Science,
Graduate School of Information Science and Engineering, Tokyo Institute of Technology
Tokyo, Japan
ananta@bi.cs.titech.ac.jp

Yutaka Akiyama
Department of Computer Science,
Graduate School of Information Science and Engineering, Tokyo Institute of Technology
Tokyo, Japan
akiyama@cs.titech.ac.jp

***Abstract*— We propose an approach for metagenome fragment binning using support vector machine (SVM) and characterization vectors. We developed this method to overcome the limitation of the composition-based approach using k-mer features to perform the binning process, particularly for short fragments. We take advantage of characterization vectors, which consider global information of DNA fragments without performing sequence alignments. The global information of sequences can be represented by twelve-dimensional information. Our experiments show that this method is highly accurate for binning metagenome fragments at the genus level with fragment lengths ≥ 500 bp for datasets representing known and new organisms. This approach is promising for extension to other taxonomy levels.**

Keywords - metagenome; binning; characterization vector; support vector machine;

I. INTRODUCTION

A common approach in studying the genetic material of any organism is to cultivate the organism in the lab and then perform shotgun sequencing and de novo sequence assembly. These processes are required to produce DNA sequences, which contain genetic information of the organism. Unfortunately, only about 1% of the many microorganisms in the world can be cultured [1]. The rest must be collected by taking samples directly from the environment. Metagenomics is the study of genetic material taken directly from the environment, which contains sequences for a mixture of genomes.

To get a complete genome from an environmental sample, the sequences obtained by shotgun sequencing must be assembled. This assembly process will yield scaffolds, contigs and unassembled reads. However, because the environmental sample contains a variety of organisms, a binning process is needed to classify them. An assembling process is required before classification to obtain contigs (long fragments with length ≥ 3 kbp) [1]. Until now, highly accurate results from the metagenomic fragment binning process were not possible using short fragments (fragment length < 3 kbp).

Composition-based binning is commonly used for the binning process. This approach uses k-mer frequencies as features. Some significant results have been obtained using this approach combined with unsupervised or supervised learning, such as PCA [1], SOM [2], k-means clustering [3], support vector machine (SVM) [4], and naïve Bayesian classifiers [5, 6, 7]. Although the use of k-mer frequencies is promising for solving classification problems of metagenomics fragments, composition-based approaches still have two limitations: (1) for fragment lengths < 6 kbp, the accuracy of the binning process decreases with decreasing DNA fragment length, and (2) the dimensionality of feature space increases drastically with large k values. PhyloPythia [4] could obtain high sensitivity (> 70%) for the genus by using fragment lengths ≥ 3 kbp; however, low accuracy resulted with fragment lengths ≤ 1 kbp.

In this study, we propose using characterization vectors reported by Liu [8] for covering the weaknesses of composition-based binning with restrictions in terms of fragment lengths and k values. Characterization vectors consist of twelve-dimensional information. These vectors have been implemented to perform clustering of DNA sequences by calculating the distance between two characterization vectors. However, this feature has not been widely used by other researchers in the field of DNA fragment clustering or classification.

II. MATERIALS AND METHODS

We used two groups of datasets generated by MetaSim [9] to represent known and new organisms. For known organisms, we generated fragments with six different lengths from ten species in three genera (Table 1). The fragment lengths were 500 bp, 1 kbp, 5 kbp, and 10 kbp, and the total fragment numbers for training and testing were 10,000 and 5,000, respectively.

For new organisms, we used nine species in three genera (Table 2) to generate fragments with four different lengths (500 bp, 1 kbp, 5 kbp, and 10 kbp). We defined these fragments for testing new organisms and used a total of 5,000 fragments.

*Also at Department of Computer Science, Bogor Agriculture University, Indonesia

Table 1 Dataset for representing known organisms

Species	Genus
Agrobacterium radiobacter K84 chromosome 2	Agrobacterium
Agrobacterium tumefaciens str. C58 chromosome circular	
Agrobacterium vitis S4 chromosome 1	
Bacillus amyloliquefaciens FZB42	Bacillus
Bacillus anthracis str. Ames Ancestor	
Bacillus cereus 03BB102	
Bacillus pseudofarmus OF4 chromosome	
Staphylococcus aureus subsp. Aureus JH1	Staphylococcus
Staphylococcus epidermidis ATCC 12228	
Staphylococcus haemolyticus JCSC1435	

Table 2 Dataset for representing new organisms

Species	Genus
Agrobacterium radiobacter K84 chromosome 1	Agrobacterium
Agrobacterium tumefaciens str. C58 chromosome linear	
Agrobacterium vitis S4 chromosome 2	
Bacillus thuringiensis str Al Hakam	Bacillus
Bacillus subtilis subsp. Subtilis str 16B	
Bacillus pumilus SAFR-032	
Staphylococcus carnosus	Staphylococcus
Staphylococcus saprophyticus subsp. Saprophyticus ATCC 1530 S	
Staphylococcus lugdunensis HKU09-01	

We designed a multiclass classifier to classify metagenomic fragments with different lengths at the rank of genus. We employed multiclass SVM classifiers trained with fragments of a certain length. The characterization vectors served as the input items for the SVM. For this purpose, we generated a Gaussian radial basis function (RBF) based on the characterization vector.

The characterization vector [8] contains twelve-dimensional information as follows:

$<n_A, T_A, D_A, n_G, T_G, D_G, n_T, T_T, D_T, n_C, T_C, D_C>$

The first four parameters, n_A, n_G, n_T, and n_C, refer to the number of A, G, T, and C nucleic bases, respectively, within the DNA sequence (Figure 1). The total number of bases represents the length of the sequence. Calculation of these four parameters is not adequate to distinguish between two DNA sequences because two DNA sequences with the same length may have the same total number of nucleic bases.

ATGCTTACGTAGCATG → n_A=4, n_G=4, n_T=5, n_C=3

Figure 1 Number of nucleic bases of sequence ATGCTTACGTAGCATG

The second four parameters, T_A, T_G, T_T, and T_C, represent the total distance of each nucleotide base to the first nucleotide. The total distance T_i is defined as:

$$T_i = \sum_{j=1}^{n_i} t_j$$

where i = A, G, T, C and t_j is the distance from the first nucleotide to the jth nucleotide specified in the DNA sequence. These distance parameters and the number of nucleic bases are used to make similarity measurements between two DNA sequences more accurate (Figure 2).

Sequence 1: ATTGCGCAAGCAG → n_T=2, T_T=5 (positions 2, 3)

Sequence 2: ACGCGCATTAAGA → n_T=2, T_T=17 (positions 8, 9)

Figure 2 the first sequence, ATTGCGAAGCAG has number of nucleic bases T (n_T) = 2 and the total distance (T_T) = 2+3 = 5. However, the second sequence ACGCGCATTAAGA has the same number of nucleic base T, (n_T) = 2, but distance (T_T) = 8+9 = 17

However, these two groups of parameters are still not sufficient to compare two DNA sequences. Figure 3 shows that two sequences can have the same number of nucleic bases T (n_T) and also the same values for the distances between the nucleic bases T, T_T. Therefore Liu et al. [8] defined a third parameter, the variance of distance, for each nucleic base as follows:

$$D_i = \sum_{j=1}^{n_i} \frac{(t_j - \mu_i)^2}{n_i}$$

where i = A, T, G, C and t_j is the distance from the first nucleotide to the jth specified nucleotide i in the DNA sequence. They defined μ_i as follows:

$$\mu_i = \frac{T_i}{n_i}$$

These three groups of parameters make similarity measurements between two DNA sequences accurate.

Sequence 1: ACTGCGTAAGCAG → n_T=2, T_T=10, D_T=4 (positions 3, 7)

Sequence 2: ACCTCTAAAGCAG → n_T=2, T_T=10, D_T=1 (positions 4, 6)

Figure 3 the first sequence ACTGCGTAAGCAG and the second sequence ACCTCTAAAGCAG has the same value of n_T=2 and T_T=10, but the first sequence has D_T=4 and the second sequence has D_T=4. Thus, parameter D makes similarity measurements more accurate.

This characterization vector contains twelve-dimensional information. We hypothesized that it could be used to characterize similarities between DNA sequences, because it describes the nucleotide contents, the distance of each

nucleic base from the first nucleotide, and the distribution of each nucleotide within the DNA sequence.

We used LIBSVM [10] to perform multiclass SVM. A 5-fold cross validation was conducted to obtain the best parameters C and γ. These two parameters were used to train the whole training set using SVM. We applied a one-versus-one technique, where classification is done by a max-wins voting strategy, in which every classifier assigns the instance to one of the two classes; then, the vote for the assigned class is increased by one vote, and finally the class with the most votes determines the instance classification. This program was run on a Dual Core AMD Opteron CPU 2.4 GHz system supplied with 32 GB of RAM.

III. Results and Discussion

A. Reason for using Characterization Vector

The objective of this paper was to perform metagenome fragment binning. The approaches used in this field include homology-based binning and composition-based binning. Homology-based binning is performed by comparing metagenome fragments against a reference database using BLAST [11] or MEGAN [12]. BLAST was developed based on nearest neighbors, whereas MEGAN applies both nearest neighbors and last common ancestors. These approaches find similarities between sequences through sequence alignment processes, which are time consuming because they consider all of the information of a sequence.

On the other hand, composition-based binning tries to perform the binning process using k-mers as features with the machine learning method. Because this approach bypasses the need for sequence alignments, it can reduce the execution time. However, this approach is restricted by the fragment lengths and k values. From PhyloPythia [4] and the naïve Bayesian classifier [5], we learned that high accuracy of the classifier can be obtained by using k=5 for long fragments (fragment length $\geq$ 3 kbp) or using k $\geq$ 9 for short fragments (fragment length = 500 bp). However, using k $\geq$ 9 for k-mers results in a high-dimension feature space (4^k). Although high accuracy can be obtained, generating feature space based on k-mers frequencies will consume much more memory and computational time. On the other hand, if small k-mers (k<5) are used, the accuracy drastically decreases. Features generated from a small k value may not be capable of capturing the main characteristics that can distinguish one taxonomic lineage from another [13].

In this paper, we used the characterization vector which considers the global information of DNA fragments without performing sequence alignments. The global information of sequences can be represented by twelve-dimensional information containing three main attributes: number of nucleic bases, the distance of each nucleic base, and the variance of distance for each nucleic base.

B. Classification results

We generated 10 kbp fragments from ten species using MetaSim for training data. These fragments were generated randomly from the whole genome sequence of each species downloaded from NCBI databases. To validate our model, we also generated 5 kbp fragments from the same genome using MetaSim to represent known organisms. We used fragment lengths of 500 bp, 1 kbp, 5 kbp, and 10 kbp.

Accuracy, sensitivity, and specificity were used to measure the performance of our classifier. The accuracy measure is defined as follows:

$$accuracy = \frac{number\ of\ correct\ predicted\ data}{total\ number\ of\ data}$$

This method can obtain high accuracy using 500 bp fragments. Figure 4 demonstrates high accuracy for the genus level. The accuracy of this method was 81% for 500 bp, 85% for 1 kbp, 90% for 5 kbp, and 92% for 10 kbp fragments, respectively, showing the potential power of the method.

This method exhibits high sensitivity with 500 bp fragments. The sensitivity of the classifier was 78% for 500 bp, 83% for 1 kbp, 89% for 5 kbp, and 92% for 10 kbp, indicating that this method can recognize at least 78% of the actual positives. In addition, the specificity of this method was between 89%-98% for all of the tested fragment lengths for the genus (Figure 4).

All of these results show that our approach using the SVM and characterization vector is promising for very sensitive and accurate classifications. Indeed, the current classifier employing SVM and k-mer features can obtain high sensitivity only with fragment lengths $\geq$ 3 kbp [4]. Moreover, the naïve Bayesian classifier and k-mer features can obtain 80% accuracy only using k=10 [5], which is also time-consuming, as this approach requires generating a high-dimension feature space (4^{10}). With our approach, we do not face a high-dimension feature space problem, because the characterization vector always contains twelve-dimensional information.

To validate our approach for the binning process for new organisms, we generated 5-kbp fragments from nine species using MetaSim. These genomes are different from the genome representing known organisms, but still belong to the same genus. The fragments lengths were 500 bp, 1 kbp, 5 kbp, and 10 kbp. Figure 5 shows that the accuracy for the genus was not as high as that for known organisms. However, the results indicate the power of our approach in classifying fragments with length < 1 kbp. The accuracy of this method was 78% for 500 bp, 80% for 1 kbp, 86% for 5 kbp, and 87% for 10 kbp.

This trend was also supported by the sensitivity and specificity. For the genus level, the sensitivity of this method reached 75%-85% for a fragment length of 500 bp to 10 kbp. Moreover, the specificity of this method for

genus increased from 86% for 500 bp fragments to 92% for 10 kbp. In general, the performance of this method slightly decreased when used to classify new organisms. The misclassification of fragments probably occurred due to sequence overlap between different strains possibly within the species in the different genus. The assignment of fragments to different genera may have been caused misclassification of fragments.

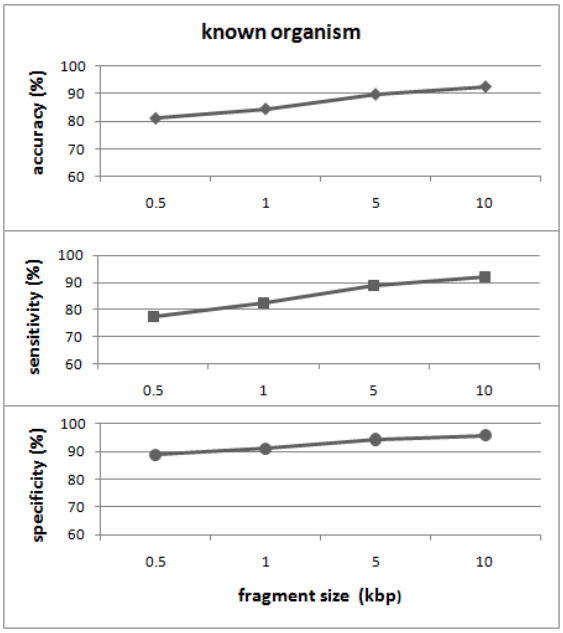


Figure 4 the performance of classifier for known organism

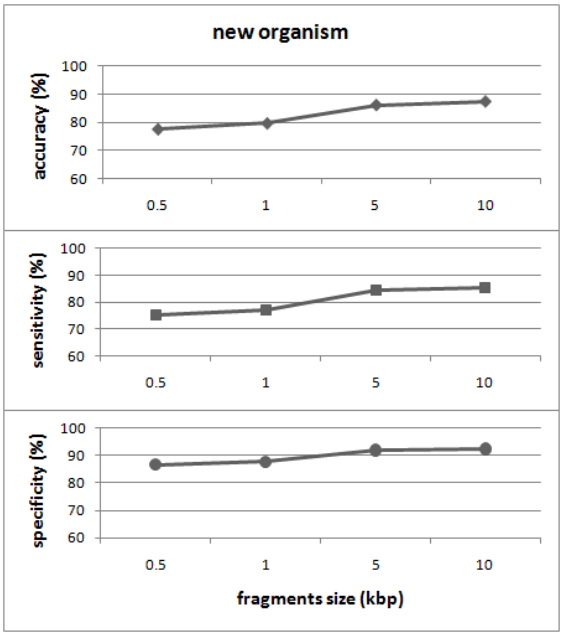


Figure 5 Performance of the classifier for new organisms

Figures 6 and 7 show the sensitivity and specificity of each genus with different fragment lengths for both known and new organisms. When using fragments belonging to species in the *Agrobacterium* genus, the sensitivity and specificity of our method was high, even using 500 bp fragments. The performance increased with increasing fragment length. For known organisms, we obtained sensitivity and specificity of 93% and 98%, respectively, with fragment lengths of 500 bp. Therefore, our method can recognize almost 93% of the fragments of species belonging to the *Agrobacterium* genus. This trend was also shown using the dataset representing new organisms.

However, the sensitivity of our method was 52% for known organisms and 40% for new organisms when classifying fragments of species in the *Staphylococcus* genus with 500 bp fragments. These results occurred because many fragments had been assigned to *Bacillus*. *Staphylococcus* and *Bacillus* are in the same order *Bacillales*. Thus, misclassification of fragments probably occurred because of sequence overlap between different strains within species in the same order. Therefore, we need to consider extending this method to classify fragments in higher taxonomy levels.

IV. Conclusion and Future Work

In this paper, we presented a metagenome fragment binning process by implementing SVM and Gaussian RBF kernels based on characterization vectors. Our results show that this approach can improve the binning process to obtain high accuracy at the genus level with short fragment lengths (500 bp). This method is promising because the characterization vector can provide enough information to perform similarity measures without doing sequence alignments or using composition measures such a k-mers.

Future work will continue in two directions: (1) to generate a new kernel by combining the characterization vectors and k-mer features and (2) to test this new method on other taxonomy levels, such as family, order, class, phylum, and domain, with more species.

Acknowledgment

This work was supported in part by the Global Center of Excellence (GCOE) CompView, Tokyo Institute of Technology.

We would like to thank Youhei Namiki for the characterization vector generation code.

References

[1] H. Wu, "PCA-based Linear Combinations of Oligonucleotide Frequencies for Metagenomic DNA Fragment Binning," Proc. IEEE Symp. Computational Intelligence in Bioinformatics and Computational Biology (CIBCB '08), IEEE Press , Sept. 2008, pp. 46 – 53, doi:10.1109/CIBCB.2008.4675758.

[2] C. K. Chan, A. L. Hsu, S. Tang, S. K. Halgamuge, "Using Growing Self –Organizing Maps to improve the binning process in environmental whole-genome shotgun sequencing," J. Biomed. and Biotech, vol. 2008, doi:10.1155/2008/513701.

[3] B. Yang, Y. Peng, H. C. M. Lung, S. M. Yiu, J. C. Chen, F. Y. L. Chin, "Unsupervised binning of environmental genomic fragments based on an error robust selection of l-mers, " Proc. Third International Workshop on Data and Text Mining in Bioinformatics (DTMBIO'09), Nov. 2009.

[4] C. McHardy, et. al.,"Accurate phylogenetic classification of variable-length DNA Fragments," Nature Methods, vol. 4 no. 1, Jan. 2007 , pp. 63-72. 2007, doi:10.1038/nmeth976.

[5] G. Rosen, , et. al., "Metagenome Fragment Classification Using N-Mer Frequency Profiles," Advance in Bioinformatics, vol. 2008, Sept. 2008, doi:10.1155/2008/205969.

[6] R. Sandberg, G. Winberg, C. I. Branden, A. Kaske, I. Ernberg, J. Coster,"Capturing whole-genome characteristics in short sequence using naïve bayessian classifier," Genome Res.,vol. 11 no. 8, Aug. 2001, pp. 1404-1409, doi: 10.1101/gr.186401.

[7] Q. Wang, G. Garrity, J. M Tiedje, J.R. Cole, "Naïve bayes classifier for rapid assignment of rRNA sequences into the bacterial taxonomy," Appl. Environ. Microbiol., vol. 73 no. 16, Aug. 2007, pp 5261-5267, doi:10.1128/AEM.00062-07.

[8] L. Liu, , et al."Clustering DNA sequences by feature vectors". Molecular Phylogenetics and Evolution", vol. 41, Oct. 2006, pp. 64-69, doi:10.1016/j.ympev.2006.05.019.

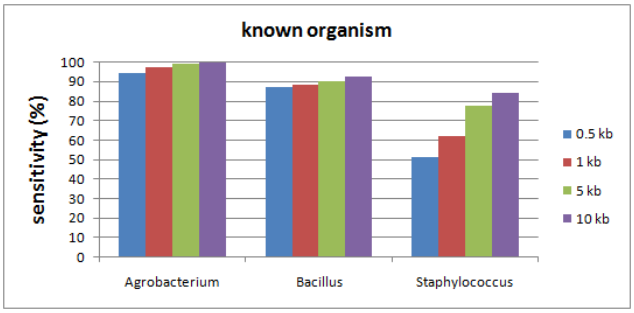
[9] Richter DC, Ott F, Auch AF, Schmid R, Huson DH. "MetaSim—A Sequencing Simulator for Genomics and Metagenomics". PLoS ONE 3(10): e3373. doi:10.1371/journal.pone.0003373. 2008.

[10] C. C. Chang and C. J. Lin." LIBSVM : a library for support vector machines". Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.

[11] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D.J. Lipman, "Basic local alignment search tool," J. Mol. Biol., vol. 215 no. 3, Oct. 1990, pp. 403-410.

[12] D. E. Huson, A. F. Auch, J. Qi, S.C. Schuster, "MEGAN analysis of metagenomic data," Genome Res., vol. 17 no. 3, Mar. 2007, pp 377-386. 10.1101/gr.5969107.

[13] K. Mavromatis, , et. al."Use of simulated data sets to evaluate the fidelity of metagenomic processing methods", Nature Methods, vol. 4 no. 6, Apr. 2007, pp. 495-500, doi:10.1038/nmeth1043.



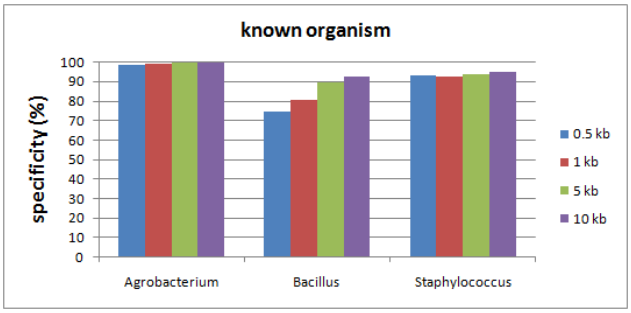
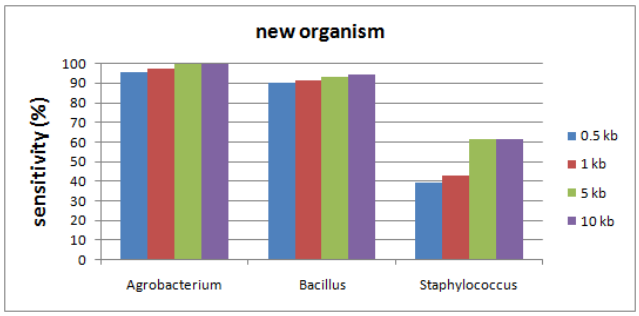


Figure 6 Sensitivity and specificity of classifiers for each genus with datasets representing known organisms. The legend specifies the color code for the respective fragment length in the graphs.



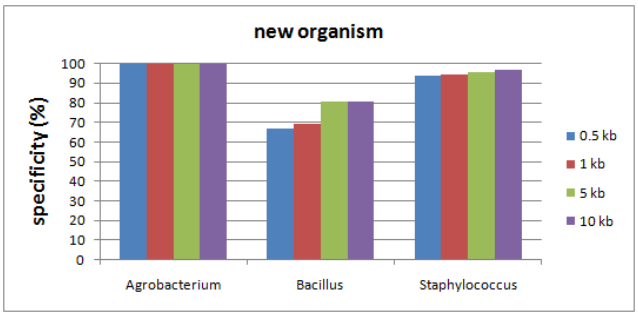


Figure 7 Sensitivity and specificity of classifiers for each genus with datasets representing new organisms. The legend specifies the color code for the respective fragment length in the graphs.