Journal of Computer Science and Cybernetics, V.30, N.3 (2014), 267–277
DOI:10.15625/1813-9663/30/3/3754

A NOVEL L-MER COUNTING METHOD FOR ABUNDANCE-BASED BINNING OF METAGENOMIC READS

LE VAN VINH[1], TRAN VAN LANG[2,3], TRAN VAN HOAI[1]

[1]*HCMC University of Technology, VNU-HCM*
[2]*Institute of Applied Mechanics and Informatics, VAST*
[3] *Lac Hong University, MOET*
langtv@vast.vn

Tóm tắt. Phân loại trình tự là bước quan trọng trong quá trình phân tích dữ liệu metagenomic. Trong khi những phương pháp không có giám sát dựa trên đặc trưng hợp thành chỉ hiệu quả cho xử lý trình tự dài, các phương pháp dựa trên độ phong phú thường được sử dụng cho phân loại trình tự ngắn. Những giải pháp phân loại dựa trên độ phong phú hiện nay thường sử dụng tần số l-mer có độ dài cố định để phân loại trình tự vào các nhóm mà các trình tự trong mỗi nhóm thuộc về các hệ gien (hay loài) có độ phong phú tương tự nhau. Tuy nhiên, hiệu năng phân loại của các giải pháp này rất nhạy cảm với độ dài các l-mer, và chúng gặp khó khăn khi phân loại những trình tự thuộc các hệ gien có độ phong phú thấp vì sự lặp lại của các đoạn l-mer ngắn trong các hệ gien này. Trong bài báo này, một phương pháp đếm mới sử dụng các l-mer có độ dài thay đổi được đề xuất, cho phép giải quyết vấn đề lặp lại của các đoạn l-mer ngắn, nhằm cải tiến độ chính xác của các giải pháp phân loại dựa trên độ phong phú. Phần thực nghiệm cho thấy rằng một giải pháp cải tiến của AbundanceBin (một phương pháp phân loại thường được sử dụng) trong đó phương pháp đề xuất được áp dụng cho độ chính xác cao hơn giải pháp ban đầu. Phần mềm hiện thực cho giải pháp này có thể được tải về tại địa chỉ: http://it.hcmute.edu.vn/bioinfo/MetaSeqBin/index.htm

Từ khóa. metagenomics, phân loại trình tự, đếm l-mer, trình tự DNA, giải mã trình tự thế hệ mới.

Abstract. The binning of reads is a crucial step in metagenomic data analysis. While unsupervised methods which are based on composition features are only efficient for long reads, genome abundance-based methods are often used in the binning of short reads. Previous abundance-based binning approaches usually use fixed-length l-mer frequencies to separate reads into groups such that reads in each group belong to genomes (or species) of very similar abundances. However, their classification performances are very sensitive to the length of l-mers, and they get difficult to separate reads from low-abundance genomes due to the repeat of short length l-mers in the genomes. In this paper, a new variable-length l-mer counting method is proposed to enable dealing with the short length l-mer repetition for improving the accuracy of abundance-based binning approaches. Computational experiments demonstrate that an improved approach of AbundanceBin (a commonly used binning method) in which the proposed method is applied achieves higher accuracy than the original one. The software implementing the approach can be downloaded at http://it.hcmute.edu.vn/bioinfo/MetaSeqBin/index.htm.

Keywords. metagenomics, binning, l-mer counting, DNA sequence, next-generation sequencing.

1. INTRODUCTION

Microbes are the most diverse forms of life on Earth and affect directly on human lives. Understanding microbial communities brings benefits to us in many fields [4]. Traditional microbial genomic studies only focus on single species in pure laboratory culture due to the limitation of experiments. However, ninety-nine percent of microbes cannot be cultured in the laboratory [1]. With the development of next-generation sequencing techniques [11], the traditional methods are gradually replaced by metagenomics. This discipline enables studying directly on genomes from an environmental sample without isolating and culturing single organism in the laboratory.

In metagenomics, data contains reads from various organisms. Thus, an important problem needed to be solved is to separate reads into groups of individual genomes or closely related organisms. It is referred to as *binning problem*. Binning approaches can be classified into two main categories: *supervised methods* and *unsupervised methods*.

Supervised methods require reference databases containing sequences with known taxonomic origin. It can be further divided into *composition-based* and *homology-based* methods. Homology-based approaches ([5, 8]) directly align DNA fragments to a reference set using an alignment tool (e.g., BLAST). In the composition-based approaches, compositional features that can be extracted from nucleotide fragments and reference sequences (e.g. oligonucleotide frequencies, GC-content) are used for classification ([2, 13]). These approaches have been shown to be accurate if there are complete reference databases. However, the ratio of microbes which have been discovered is very small [1]. In addition, a recent study [3] showed that only 12% real metagenomic data from a coral atoll can hit against the 16S rDNA database by BLAST tool.

Unsupervised methods can be used to overcome the lack of reference data sets. The methods rely on features extracted from data samples. Some approaches are based on compositional features such as LikelyBin [7], Scimm [6], MetaCluster 1.0, 2.0 ([19, 20]). While LikelyBin and Scimm use Markov chain model for classification, MetaCluster 1.0 and 2.0 are k-means based methods. The approaches are only efficient for long reads ($>$ 1kbp), and get low performance for short reads (50-400bp). Moreover, their performances drop significantly when the abundance level of genomes is very different [20].

Some other unsupervised approaches are based on the abundance level of genomes such as MetaCluster 5.0 [17], AbundanceBin [18], and Olga *et al* [16]. MetaCluster 5.0 separates reads into three groups with different abundance levels (high, low and extremely low level) and applies different further classification strategies for each group. AbundanceBin and Olga *et al.* are two recent approaches for binning reads only relying on genome abundances. The approaches aim to group reads which belong to genomes of similar abundance levels. In AbundanceBin, the number of occurrences of l-mers (with a sufficient value of l) are assumed to follow a Poisson distribution. An expectation maximization algorithm is used to estimate genome abundances and genome sizes. Olga's approach uses a similar method with AbundanceBin, but the method improves a task of l-mer counting by applying an idea from a Balls and Bins problem to deal with sequencing errors. The two approaches can be used as a preprocessing task of other binning approaches for performance improvement.

The abundance-based binning methods are shown to be able to classify short reads. However, due to the repeat of short length l-mers in genomes, previous approaches get difficult to accurately estimate the genomes abundances as well as separate short reads. Dealing with the problem, this work focuses on the task of l-mer counting of a binning method. A novel l-mer counting method proposed in this paper aims to better reflect the abundance level of genomes for increasing classification performance. Some l-mer counting methods were proposed in recent years ([10, 15]). However, they

only focus on storage space and computation time, not strongly focusing on the quality of the binning process.

The following sections of this paper are organized as follows: In section II, the abundance-based binning problem is stated, then the problem of l-mer length and a proposed l-mer counting method are presented. Section III shows experimental results. The last section provides a conclusion and future works.

2. METHODS

2.1. Problem statement

In this section, the problem of abundance-based binning for metagenomic reads using l-mer frequencies is established. An assumption used for the problem is that the l-mer frequency from reads of a genome is proportional to that of genome abundance ([17, 18]). A major objective of the problem is to classify l-mers into bins (or clusters) using their frequencies, and then reads are assigned to bins basing on the membership degrees of their l-mers to the bins. The sub-problem of l-mers binning can be formulated as follows:

Given a set of data containing reads from the genomes with k different abundance levels. Let $L = \{l_1, l_2, ..., l_m\}$ be a set of distinct l-mers (m be the number of distinct l-mers), $C = \{c_1, c_2, .., c_m\}$ be a set of distinct l-mer counts. Let a decision variable x_{ij} be the membership degree of l-mer l_i in bin j, $1 \leq i \leq m, 1 \leq j \leq k$ and $0 \leq x_{ij} \leq 1$.

An objective function, $f(C)$, should be minimized or maximized under the constraint on the sum of membership degree of l-mers in bins as follows:

$$\sum_{j=1}^{k} x_{ij} = 1, 1 \leq i \leq m. \tag{1}$$

2.2. The problem of l-mer length

Estimation of l-mer frequencies directly affects the performance of a binning approach. However, choosing a value of l such that l-mer frequencies accurately reflect the abundance of genomes is difficult. If l is not large enough, many l-mers in genomes are repeated. Because it is not known an l-mer is repeated by others in genomes or in overlapping regions between reads, the number of occurrences of the l-mers may reflect wrong the genome abundances. Conversely, if l is too large, most of the l-mers in genomes are unique. However, because of the short overlapping regions between reads, the number of occurrences of l-mers in a set of metagenomic reads can be low. It leads to the difficulty of separating the reads, especially for short reads and low-abundance genomes. Both cases make low classification performance of abundance-based binning approaches.

For instance, in Figure 1, R1 and R2 are two DNA regions in two genomes (or in the same genome). Assuming the abundance level of the genomes is 3. Let $l = 5$. Because 5-mer u and v are the repeat of each other, u's count (or v's count) in a set of reads from the genomes is much higher than the genomes abundance (6 compare with 3). When the value of l has increased to 6, most of 6-mers are unique in the genomes (for examples w, y). However, their counts are lower than 3 in the set of reads. The count of w is 2, and the count of y is 1.

A test case is conducted for observing the average ratio of repeated l-mers to total of l-mers in genomes. Data of the test includes 30 pairs of genomes from different species in the same genus (genus

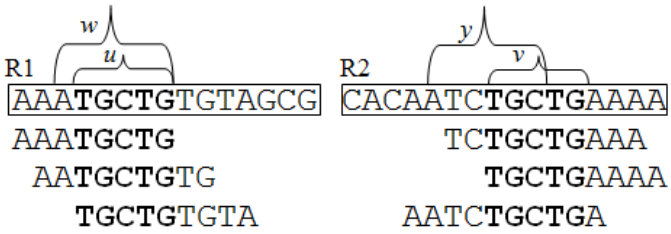

Figure 1: R1 and R2 are two DNA regions in two genomes (or in the same genome). u and v are two 5-mers, and they are repeated of each other in the genomes. w and y are two 6-mers

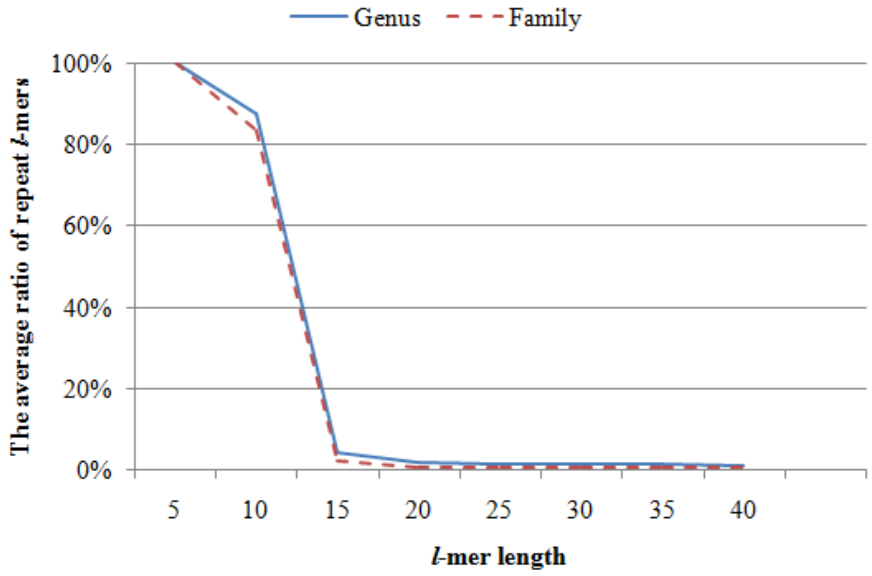

Figure 2: The average ratio of repeat l-mers with different values of l

level), and 30 pairs of genomes from different genus in the same family (family level). Figure 2 presents the results of the test. If $l < 15$, the percentage of l-mer repeats is very high. When $l > 20$, the number of l-mer repeats is lower than 4% for both genus level and family level. In another test, data containing 150bp reads from one genome (*Candidatus Blochmannia floridanus* in this test) with abundance of 10 was generated. Choosing $l = 20$, there is only 37.84% of 20-mers which their counts are ≥ 10. Thus, there is no any suitable values of l which can adapt the two aspects of the problem. However, this problem has not been solved. In other research, the value of l is chosen relatively large (20bp in [18]), or 16bp in [17]). (In [16], the authors did not mention the length of l-mers).

2.3. Proposed counting method of l-mers

In this section, l-mer counting method is proposed to address two aspects of the problem. The method aims to use large enough l value such that most of l-mers are unique in genomes, and it computes l-mer frequencies based on a small value for more accurate reflecting abundance of the genomes.

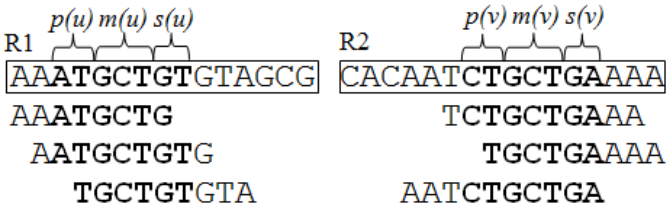

Figure 3: l-mer $u = \langle p(u), m(u), s(u)\rangle$ and $v = \langle p(v), m(v), s(v)\rangle$ in two regions R1 and R2

A variable-length l-mer (with the maximum length of l) is defined as a set of three parts, including: pre-l-mer, main-l-mer, and suf-l-mer. A main-l-mer is the middle part of l-mer, and its length is fixed to a given number of l_m. A pre-l-mer and a suf-l-mer are the two remaining parts, locating in the beginning and at the end of l-mer respectively. Their lengths are varied and limited by given numbers l_p and l_s respectively ($l = l_p + l_m + l_s$).

From the definition, this method uses a value of l_m to compute frequencies of l-mers, and a value of l to check whether the l-mers are repeated in genomes or not. Therefore, l_m should be chosen relatively small and l be large enough such that most of the l-mers are unique in genomes. For more details, the method computes the number of occurrences of main-l-mers in a set of metagenomic reads. Because of the small length, repeats of the main-l-mers may exist in genomes. To distinguish whether two main-l-mer repeats are from the overlapping regions between reads or not, our method compares their pre-l-mers and their suf-l-mers. If the pre-l-mers or the suf-l-mers are different from the others, two l-mers containing the two main-l-mers should be counted as two distinct ones. The comparison method of two l-mers is presented as follows.

Let $u = \langle p(u), m(u), s(u)\rangle$ be an l-mer. $p(u), m(u)$ and $s(u)$ be pre-l-mer, main-l-mer, and suf-l-mer of l-mer u respectively. They are strings of characters in the set{A, C, G, T}. Let $|p(u)|$, $|s(u)|$ be the lengths of $p(u), s(u)$ respectively ($|p(u)| \leq l_p, |s(u)| \leq l_s$). Let $v = \langle p(v), m(v), s(v)\rangle$ be another l-mer. Denote $g(s, pos, len)$ be a function to get a copy of a substring of string s which starts at position pos and spans len characters. $u = v$ if and only if:

$$\begin{cases} m(u) = m(v) \text{ and} \\ s(u) = g(s(v), 1, |s(u)|), \text{ if } |s(u)| \leq |s(v)| \text{ and} \\ s(v) = g(s(u), 1, |s(v)|), \text{ if } |s(v)| < |s(u)| \text{ and} \\ p(u) = g(p(v), |p(v)| - |p(u)| + 1, |p(u)|), \text{ if } |p(u)| \leq |p(v)| \text{ and} \\ p(v) = g(p(u), |p(u)| - |p(v)| + 1, |p(v)|), \text{ if } |p(v)| < |p(u)|. \end{cases} \tag{2}$$

For example, we use this method to solve the problem in Figure 1. Choose l be 7, l_m be 3 and maximum lengths of the remaining sub-parts be: $l_p = 2$ and $l_s = 2$. Assume that there are two 7-mers $u = \langle p(u), m(u), s(u)\rangle$ and $v = \langle p(v), m(v), s(v)\rangle$ (Figure 3). They are in the repeated regions. Thus, the count of u's main-l-mer (or v's main-l-mer) in the set of reads is 6. After their pre-l-mers and suf-l-mers are compared, they are regarded as two distinct l-mers with the count of 3. The repeats of variable-length 7-mers u in the set of reads are ATGCTG, ATGCTGT and TGCTGT, while the repeats of variable-length 7-mers v are CTGCTGA, TGCTGA and CTGCTGA.

2.4. Binning approach

The l-mer counting method is applied in a proposed binning approach for enhancement of classification performance. This approach can be known as an improvement of AbundanceBin - a commonly used abundance-based binning approach. The proposed algorithm includes three steps. In the first step, counts of l-mers are computed by the proposed counting method. The second step separates the l-mers into bins using their frequencies. Reads are assigned to the bins in the final step. Methods used in the last two steps are same as those of AbundanceBin.

2.4.1. Step 1: Counting l-mers

In this step, the number of occurrences of l-mers is computed by the proposed method. The values of l_m, l_p, and l_s have to be given. A hash table is used to store l-mers for reduction searching time. The hash size can be changed for efficient memory usage.

2.4.2. Step 2: Binning of l-mers

In the proposed approach, using the same ideas with AbundanceBin [18] and Olga *et al.* [16], reads sampled from the genomes are assumed following the Poisson distribution [9]. Let g_j be a genome, with a length of $|g_j|$. Let l_i (with a length of l) be an l-mer coming from the genome. Let $|r|$ be the length of a read. If the number of reads generated from genome g_j is n_j, the number of occurrences of an l-mer l_i should follow a Poisson distribution with parameter $\lambda_j = n_j(|r| - l + 1)/(|g_j| - |r| + 1)$. From the assumption, the objective function $f(C)$ in section of the problem statement is presented in the probabilistic manner as follows:

$$\hat{\theta}_{ML} = arg\max_{\theta}\{\log p(X|\theta)\} \tag{3}$$

in which $X = \{c_i\}$, $\theta = \{(g_j, \lambda_j), 1 \leq j \leq k\}$.

An Expectation Maximization (EM) algorithm is used to approximate the abundance level and the size of each genome. It is an iterative method for finding the maximum likelihood estimate of the parameter θ (noted by $\hat{\theta}_{ML}$) in the model. After the parameter θ is assigned initial values, the following two tasks are repeated until the parameter converges or the number of iterations exceeds a given threshold: First, calculate the membership degree of l-mers in bins given c_i and θ. Second, update the parameter θ for the next iteration given the membership degree and c_i.

2.4.3. Step 3: Assigning reads to bins

In this step, the reads are assigned to the bins based on the results of their l-mer binning. A read is assigned to a bin if the product of the membership degrees of its l-mers in the bin is the highest value among all bins. A read is not assigned to any bins if the highest membership degree is smaller than 50%.

The proposed algorithm uses the same technique as those of AbundanceBin to deal with sequencing errors and extremely high l-mer counts. Two thresholds $count_{min}$ and $count_{max}$ are used to limit unexpected l-mers counts. A l-mer count is used for abundance level estimation task must satisfy the following condition: $count_{min} \leq l$-mer count $\leq count_{max}$. When more than 90% of l-mers of a read are excluded by this technique, the read is not assigned to any bins.

3. EXPERIMENTAL RESULTS

The performance of the proposed approach is evaluated on simulated data, downloaded from the National Center for Biotechnology Information (NCBI) database. A commonly used tool, MetaSim [14], is used to generate data. We create reads from bacterial genomes for both cases of with and without sequencing errors. Error-free sequencing reads are created by the exact simulator setting of MetaSim, while error sequencing reads follow the Roche 454 sequencing model. The read length is from 75bp to 400bp.

The proposed approach is compared with AbundanceBin in two experimental scenarios. The first scenario is to evaluate the two approaches in separating reads into bins, such that each bin contains reads having very similar abundance levels. In the second scenario, each approach is combined with MetaCluster 2.0 to classify reads into different groups of species. Note that we do not combine the abundance-based methods with the higher versions of MetaCluster (3.0, 4.0, and 5.0) because the feature of genome abundance is used in the approaches. Thus, it is not suitable for this situation. Two statistical measures, including *sensitivity* and *accuracy* defined as in [12] are used for evaluation.

3.1. First scenario - separating reads into groups of different genome abundances

The observation of data (Fig. 2) shows that the number of occurrences of l-mers are very high if $l \leq 15$ ($> 15\%$). This explains why the performance of AbundanceBin is very low with the value of l, and it achieves the best performances with $l \geq 20$ in most of cases [18]. In this test case, the proposed approach is compared with AbundanceBin with $l = 20, 25$, and 35. In the proposed approach, $l_m \leq 15$ and $l_p = l_s = (l - l_m)/2$ for all cases are chosen. Therefore, groups of three values (l_p, l_m, l_s) corresponding to the lengths of $l = 20, 25$, and 35 used in the proposed approach are $(4, 12, 4), (5, 15, 5)$, and $(10, 15, 10)$ respectively. Besides, in cases of sequencing error, the same value of $count_{min}$ and $count_{max}$ for both approaches was set.

Table 1 shows results of 7 data sets(denoted by from T1 to T7). The proposed algorithm has better performance (about $0.07\% - 5\%$) comparing with AbundanceBin in most of the test cases. As we can see in the table, if read lengths decrease, the improvement slightly increases. For instance, the improvement for data setswith read lengths of 400bp is about 0.2% to 1%, while the improvement is about 0.07% to 5% for data setswith read lengths of 150 bp and 75 bp. In case of the test on data set T6, AbundanceBin fails to separate reads if $l = 35$, whereas our approach can classify reads into 3 clusters. Results on samples from T1 to T6 show the strength of our approach comparing with AbundanceBin when they work with reads from the genomes of small different abundances. Results on sample T7 present a fact that both approaches work quite efficiently on reads from genomes of large different abundances. Note that when $l_p = l_s = 0$, the proposed approach returns the same results with AbundanceBin for all cases. Besides, because the latest version of AbundanceBin only runs with $l \leq 28$ (by our experiments), we use the results of the proposed approach with $l_p = l_s = 0, l_m = 35$ as the results of AbundanceBin with $l = 35$.

For more comparison, we deploy another test case to evaluate both of approaches with various length of l-mers. A set of error-free reads is generated from two genomes: *Acetohalobium arabaticum DSM 5501* and *Acidianus hospitalis W1* with the abundance level of 5 and 10 respectively. The read length is 150bp. Fig. 4 presents the performances of the two approaches with the values of l from 10 to 60. For the proposed approach, when $l > 15$, l_m is always set by 15. The proposed approach has better results (sensitivity value) than AbundanceBin in all cases of l. With $l < 20$, the performance

ID	Abundance level	Read length (bp)	l	AbundanceBin Sen. (Acc.)	Proposed App Sen. (Acc.)
T1	5	400	20	82.69%(82.69%)	**83.21%(83.21%)**
	10		25	83.39%(83.39%)	**83.92% (83.92%)**
			35	83.79%(83.79%)	**84.8% (84.8%)**
T2	4	400	20	85.56%(90.37%)	**85.71% (90.47%)**
	8		25	85.77%(90.51%)	**86% (90.68%)**
	20		35	85.88%(90.58%)	**86.26% (90.84%)**
T3*	6	150	20	92.56%(92.56%)	**94.96% (94.96%)**
	18		25	89.69%(89.69%)	**90.58% (90.58%)**
			35	64.3%(64.3%)	**67.12% (67.12%)**
T4	5	150	20	81.7%(81.7%)	**83.3% (83.3%)**
	10		25	81.78%(81.78%)	**83.61% (83.61%)**
			35	81%(81%)	**84.45% (84.45%)**
T5	3	75	20	**73.89% (73.89%)**	73.67%(73.67%)
	9		25	**74.08% (74.08%)**	73.76%(73.76%)
			35	**74.37% (74.37%)**	73.81%(73.81%)
T6*	2	75	20	50.49%(75.74%)	**56.55% (78.29%)**
	4		25	46.57%(73.28%)	**55.73% (77.86%)**
	8		35	-	**50.5% (67.2%)**
T7	18	150	20	99.85%(99.85%)	**99.89% (99.89%)**
	3		25	99.81%(99.81%)	**99.88% (99.88%)**
			35	99.47%(99.47%)	**99.88% (99.88%)**

Table 1: The performances of AbundanceBin and the proposed approach. Test T3* and T6* are for sequencing error reads. Symbol "-" means binning approaches fail to separate reads. Two statistical measures used are *sensitivity* (Sen.) and *accuracy* (Acc.).

of the two approaches drops significantly. The sensitivity of AbundanceBin reaches the highest values with $20 \leq l \leq 30$, and its performance declines considerably when $l > 30$. The proposed method also gets high sensitivity value with $l \geq 20$. However, in contrast with AbundanceBin, since the value of l_m is always 15, the proposed approach returns slightly better results when the length of l increases.

3.2. Second scenario - combining with MetaCluster 2.0

Some composition-based methods (e.g., MetaCluster 2.0) use species-specific features to classify reads into different groups of species. However, their performances are low when the abundance levels of genomes are very different [20]. An abundance-based binning approach can be used as a preprocessing task of composition-based approaches for performance improvement.

In this test scenario, the performances of the two approaches are compared when combining with MetaCluster 2.0 to separate reads. Six test data are generated, including different number of genomes and abundance ratios. Reads follow 454 sequencing model with the length of 150bp. The value of l used in AbundanceBin is 25, while the values of (l_p, l_m, l_s) used in the proposed approach are $(5, 15, 5)$.

The results in table 2 show that when the abundance levels of genomes are different, the combi-

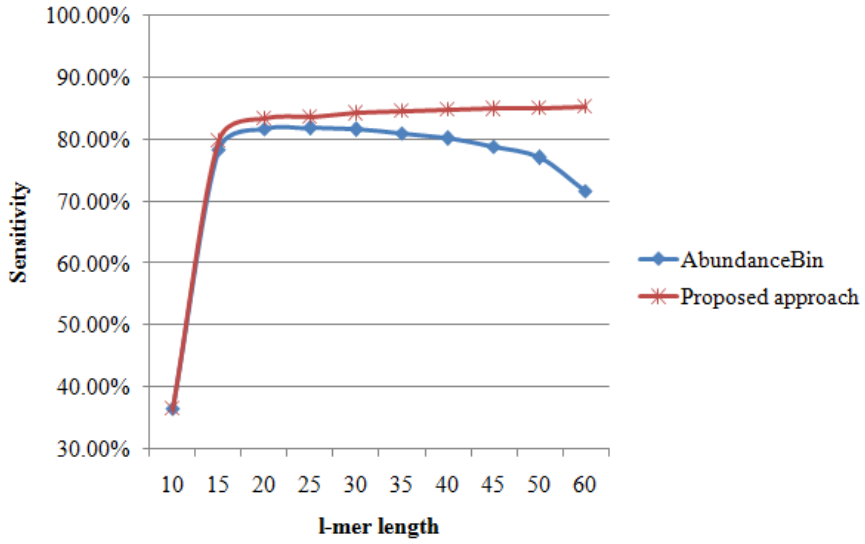

Figure 4: The performance of AbundanceBin and the proposed approach with different values of l.

ID	No. of genomes	Abundance level	MC 2.0 Sen. (Acc.)	MC 2.0 + AbundanceBin Sen. (Acc.)	MC 2.0 + Proposed App Sen. (Acc.)
S1	3	5:5:15	61.67% (74.44%)	95.14% (96.76%)	**96.31%** **(98.15%)**
S2	3	5:15:15	61.14% (74.09%)	91.5% (94.33%)	**92.42%** **(94.94%)**
S3	3	5:7:7	**74.91%** **(83.27%)**	- -	- -
S4	4	5:5:15:15	44.23% (72.11%)	88.54% (94.27%)	**89.67%** **(94.83%)**
S5	4	5:5:10:15	47.0% (73.5%)	46.65% (73.32%)	**47.92%** **(73.96%)**
S6	5	5:5:20:20:30	24.3% (69.72%)	55.8% (83.51%)	**60.88%** **(89.79%)**

Table 2: Combining AbundanceBin and the proposed method with MetaCluster 2.0. Symbol "-" means the binning approaches fail to separate reads. Two statistical measures used are *sensitivity* (Sen.) and *accuracy* (Acc.).

nation of the proposed approach with MetaCluster 2.0 can achieve better performances than MetaCluster 2.0 and the combination of AbundanceBin with MetaCluster 2.0 as well. However, in the test case S3, because the abundance levels of genomes are relatively similar (5:7:7), AbudanceBin and the proposed approach fail to separate reads. Besides, in the test case S5 (abundance levels are 5:5:10:15), the combination of AbundanceBin and MetaCluster 2.0 returns worse result than MetaCluster 2.0 and the combination of the proposed approach with MetaCluster 2.0.

In this research, the aspects of computational time and memory cost are still not considered.

Some methods in [10, 15] can be applied to decrease searching time and storage space for the two approaches.

4. CONCLUSIONS

The complexity of microbial communities requires efficient tools to analyze their data samples. Thus, different aspects of data features must be well considered. Using the feature of genome abundance can help us get better classification of metagenomic reads, especially for short reads. This paper proposes a novel l-mer counting method which is helpful to improve an abundance-based binning approach. The improved approach of AbundanceBin proposed in this paper can achieve higher accuracy comparing to the original one. This method can be used to improve other abundance-based binning approach using l-mer counts. In future works, our concerned research topic could be an automatic detection of number of bins, and an application of the proposed method to other composition-based binning approach.

ACKNOWLEDGMENT

The authors would like to thank HCMC University of Technology, Viet Nam National University Ho Chi Minh city for supporting this study. The experiments presented in this paper are tested in the High Performance Computing Center (HPCC) of the faculty of Computer Science and Engineering, HCMC University of Technology, Vietnam. We also thank Prof. Francis Chin and Dr. Yi Wang of the University of Hong Kong for their help with our research.

REFERENCES

[1] R. I. Amann, W. Ludwig, and K. H. Schleifer, "Phylogenetic identification and in situ detection of individual microbial cells without cultivation," *Microbiol Rev.*, vol. 59, no. 1, pp. 143–169, March 1995.

[2] N. N. Diaz, L. Krause, A. Goesmann, K. Niehaus, and T. W. Nattkemper, "Tacoa: Taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach," *BMC Bioinformatics*, vol. 10, no. 1, p. 56, 2009.

[3] E. A. Dinsdale, O. Pantos, S. Smriga, R. A. Edwards, and F. Rohwer, "Microbial ecology of four coral atolls in the northern line islands," *PloS One*, vol. 3, no. 2, p. e1584, 2008.

[4] J. Handelsman, "The new science of metagenomics - revealing the secrets of our microbial planet," National Academy of Sciences, USA, Tech. Rep., 2007.

[5] D. H. Huson, "Megan analysis of metagenomic data," *Genome Research*, vol. 17, no. 3, pp. 377 – 386, March 2007.

[6] D. R. Kelley and S. L. Salzberg, "Clustering metagenomic sequences with interpolated markov models," *BMC Bioinformatics*, vol. 11, no. 544, pp. 25 – 36, December 2010.

[7] A. Kislyuk, S. Bhatnagar, J. Dushoff, and J. S. Weitz, "Unsupervised statistical clustering of environmental shotgun sequences," *BMC Bioinformatics*, vol. 10, no. 316, pp. 25 – 36, December 2009.

[8] L. Krause, "Phylogenetic classification of short environmental dna fragments," *Nucleic Acids Research*, vol. 36, no. 7, pp. 2230–2239, February 2008.

[9] E. S. Lander and M. S. Waterman, "Genomic mapping by fingerprinting random clones: a mathematic alanalysis," *Genomic*, vol. 2, no. 3, pp. 231 – 239, 1988.

[10] P. Melsted and J. K. Pritchard, "Efficient counting of k-mers in dna sequences using a bloom filter," *BMC Bioinformatics*, vol. 12, no. 333, pp. 652 – 653, August 2011.

[11] M. L. Metzker, "Sequencing technologies - the next generation," *Nature Review*, vol. 11, no. 1, pp. 31 – 46, December 2010.

[12] D. L. Olson and D. Delen, *Advanced Data Mining Techniques*, 1st ed. US: Springer, January 2008.

[13] Z. Rasheed and H. Rangwala, "Tac-elm: Metagenomic taxonomic classification with extreme learning machines," *Journal of Bioinformatics and Computational Biology*, vol. 10, no. 5, October 2012.

[14] D. C. Richter, F. Ott, A. F. Auch, R. Schmid, and D. H. Huson, "Metasim - a sequencing simulator for genomics and metagenomics," *PLoS ONE*, vol. 3, no. 10, pp. 652 – 653, 2008.

[15] G. Rizk, D. Lavenier, and R. Chikhi, "Dsk: k-mer counting with very low memory usage," *Bioinformatics*, vol. 29, no. 5, 2013.

[16] O. Tanaseichuk, J. Borneman, and T. Jiang, "A probabilistic approach to accurate abundance-based binning of metagenomic reads," *Algorithms in Bioinformatics*, vol. 7534, 2012.

[17] Y. Wang, H. C. Leung, S. M. Yiu, and F. Y. Chin, "Metacluster 5.0: a two-round binning approach for metagenomic data for low-abundance species in a noisy sample," *Bioinformatics*, vol. 28, no. 18, pp. i356 – i362, September 2012.

[18] Y. W. Wu and Y. Ye, "A novel abundance-based algorithm for binning metagenomic sequences using l-tuples," *Journal of Computational Biology*, vol. 18, no. 3, pp. 523 – 534, 2011.

[19] B. Yang, Y. Peng, H. C. M. Leung, S. M. Yiu, J. C. Chen, and F. Y. L. Chin, "Unsupervised binning of environmental genomic fragments based on an error robust selection of l-mers," *BMC Bioionformatics*, vol. 11, 2010.

[20] B. Yang, Y. Peng, J. Qin, and F. Y. L. Chin, "Metacluster: unsupervised binning of environmental genomic fragments and taxonomic annotation," in *ACM BCB'10*.

Received on October 14 - 2013
Revised on January 14 - 2014