Kỷ yếu Hội nghị Quốc gia lần thứ X về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR); Đà Nẵng, ngày 17-18/08/2017
DOI: 10.15625/vap.2017.00094

PROBLEMS OF MINIMUM SIZE TO CLUSTER METAGENOMIC DATA

Van Dinh Vy Phuong[1,3], Tran Van Lang[3], Tran Van Hoai[1], Le Van Vinh[2]

[1] Faculty of Computer Science and Engineering, Ho Chi Minh City University of Technology

[2] Faculty of Information Technology, HCMC University of Technology and Education

[3] Faculty of Information Technology, Lac Hong University

phuongvdv@cse.hcmut.edu.vn, lang@lhu.edu.vn, hoai@cse.hcmut.edu.vn, vinhlv@fit.hcmute.edu.vn

ABSTRACT: *The paper review methods to binning metagenomic, such as: use k-mer to find the features, use k-mer to create a document to find hidden models, then groups sequence base on this models. To increase performance, mostly reduce the size of original data, binning directly from representation sequences. There are problems when reducing the size and only find feature to grouping from seed sequences. This paper presents an idea, using minimum vertex vertices or k-mer sets to solve the problems.*

Keywords*: metagenomic, binning, sequence, k-mer.*

INTRODUCTION

Today, Bioinformatics is not strange to who research about biological data. It is an interdisciplinary field, develops methods between software tools and biological to understand data. In 1965, Margaret Dayhoff was a pioneer in using the computer to understand biochemical ; Ben Hesper coined from 1970, with many points in the last, the mean of bioinformatics now combining BIO from biochemical and pharmaceutical data and INFORMATICS from computer science [1], otherwise, are combining to statistic, mathematics and engineering. Although Bioinformatics was concerned from 70's, over decades, it continues has a lot of challenge. Specifically, the Information Science and Technology Initiative Consortium of the National Institutes of Health in 2000, Dr. Michael Huerta redefined Bioinformatics and Computational Biology - « the Research, development or application of computational tools and approaches to leverage maximum biological, medical, behavioral or health, including those tools that allow us to acquire, store, organize, archive, analyze, or visualize such data » [1]. That mean, in the 21st century, there are a lot thing reveal the hidden diversity of DNA in the environmental genomics to understand the living world still working. As know, metagenomics (or environmental genomics, ecogenomics, community genomics) – a new method called Next Generation Sequencing to study genetic material directly from the environmental sample. It helps reduce time to get the result tests and cost.

There are a number of methods offered in metagenomic sequence analysis. The identification and classification of current sequences are based on such characteristics as homology-based, composition-based. The homologous method has the advantage of being highly accurate if the sequences need to be analyzed in a similar way or in close proximity to the sequences already in the database. This method has disadvantages depending on the source of reference data since more than 90% of the current data does not have accurate data for reference. A compositional approach that implements character-based sequencing is derived directly from the components of the metagenomic sequence. At present, the method based on composition is divided into three groups: supervised, unsupervised and semi-supervised. However, there is no solution that is considered optimal and most accurate, resolved to the individual to the individual.

In section 2, review the algorithm base on feature k-mer [2], hidden model [3, 4, 5, 6]; identify the problems in this methods. In section 3, present the methods can solve the problems in section 2. Section 4 present result of experiments and conclusion.

REVIEW METHODS

[2] Proposed a two-phase classification algorithm with two phase called BiMeta to binning group sequence in metagenomic using the l-mer frequency on non-overlapping. The main idea of [2], instead of finding the k-mer characteristic of all data in metagenomic, look for features based on sequences which do not overlap information. In which, the overlap sequences have exactly the same l-mer segment (where, l is long enough - about 20 or more and q is an estimate, in [2] q = 5). The first phase of [2] defines groups contents sequence overlap l-mer q times and identifies seed groups as non-overlapping sequences, shown in **Figure *1***. The second phase combine groups in phase 1 based on l-mer frequency distribution of non-overlapping sequences using k-means algorithms. The experimental results comparing BiMeta, MetaCluster (5.0) and AbundanceBin show an improvement of BiMeta accuracy [2]. The accuracy of phase 1 building the seed and group depends on l-mer, number threshold q to finding duplicate sequences as well as increasing the accuracy of. How many for l in i-mer and q's threshold are correct? Did may lost features from other sequences in groups if only find the base on the sequences in seed group of phase 1? Problems have to prove and have many other experiments to check that.

[5] Change the algorithm is used in phase 2 of [2] to evaluate the loss of characteristic information when chooses the seed nodes in groups. [5] applying the Latent Dirichlet Allocation (LDA) to find hidden models [7] to assess whether metagenomic's feature information is missing when [2] analyze only on the seed sequences, output of phase 1. [5] performs through 5 steps, including step 1, finding the seed sequence - is the first phase in [2]. Step 2, from the seed sequences, represent for each group, proceeds translate these sequences into the documents, which to find the hidden topic by algorithm [7] in step 3. [5] select k = 4 to create words in the vocabulary, used to find hidden models. This is a problem too and will focus in section 3. Step 4 find out the distribution of topic, word-topic, topic-document in documents. Step 5 is the final step in [5], group documents or sequences to groups base on distributions. Model of LDA in **Figure *2***. The result of [5] show that, though the disparity rate of groups in [5] is smaller than [2], the accurate is smaller (experiment results in section 4). Two problems can infer from these results: The first is the data are lost when only translate seed to document. The second is choose k - the length of words in the vocabulary.

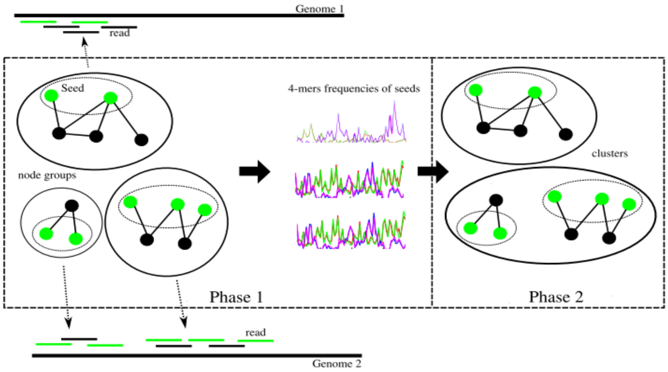


Figure 1. Two-phase of Bimeta Algorithm

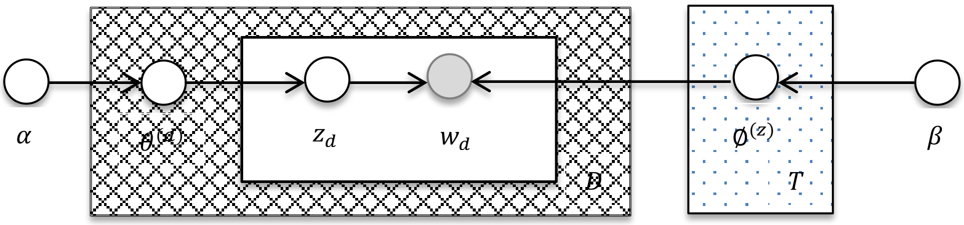


Figure 2. Latent Dirichlet Allocation

z_d: Topics in document d.
D: Documents (translate from sequences) to find hidden topics. $d \subset D$.
N_d: Number of words in document d.
w_d: Set of words in d.
T: The number of topics.
$\emptyset^{(z)}$: Distribution of word to topic z.
$\theta^{(d)}$: Distribution of topic in document d.

METHODS TO SOLVE PROBLEMS

The Vertex Cover Algorithm

Inferences from section 2, [2, 5] show that draw the features only from seed sequences make the progress more quickly but the height of ability lose the data in metagenomic. The idea to solve this problem is instead of choosing seed sequences - non-overlap sequences in the group, choose the sequences to cover all information, all sequences in metagenomic (cover all the links between sequences). **Figure *3*** presents the idea of this method. Left of **Figure *3*** is the way to choose seed sequences (seed nodes – Black/Red node). Easily remark that one information link is lost. The right of **Figure *3*** shows the other way to chose seed nodes and deduce all data. This method use algorithm called The Vertex Cover, mean minimum vertex cover. The advantage of this methods is not to scan all data in metagenomic, but still, maintain hidden features to get in later.

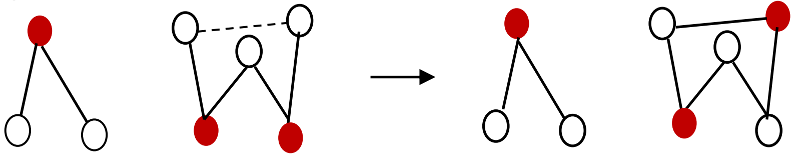


Figure 3. Seed node in Bimeta (left) and Seed nodes in Vertex Cover (right)

Minimum vertex cover on a graph is one of twenty-one NP-complete problems, which was introduced by R. M. Karp in Reducibility among combinatorial problems, 1972. Given a graph, one must find the smallest set of vertices such that every edge has at least one end vertex in the set. The set of minimum vertex cover of the general graph can be very difficult to find. 2006, Ashay Dharwadker present in [8] a method with polynomial-time to find the minimal vertex and cover all edge in graphs called Vertex Cover Algorithm.

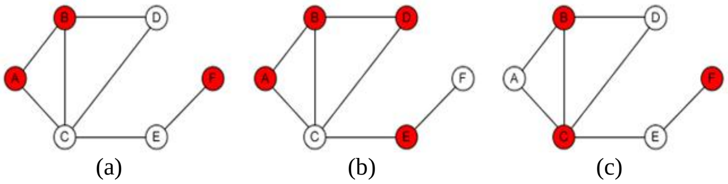


Figure 4. Minimum vertex cover

In Figure 3, image (a) is not a vertex cover. Image (b) and (c) are vertex cover, but image (c) has miminum vertex than (b).

Definitions and terms

[8] Definitions a graph G with G = (V, E). Set V is n vertices, |V| = n and set E edges link between vertices in G, each edge is an unweight and unordered pair connect vertices in G (can not full graph). The graph is not looped and not multiple edges (in this case). V and E are finite. In this paper, call 1, 2,.. n is the name of vertices in G. {u, v} is edge in G, uv ∈ E, u is neighbor of v. The degree of vertex v denoted by d(v), that is the number of neighbors of v. The maximum degree vertex in G is denoted by Δ. To identify the adjacency of the vertices, one matrix with a value of row u and column v is 1 to know that vertex u and v connect in a graph or uv ∈ E, otherwise is 0. Q is clique the graph of G if every edge connected by vertices in Q $\in E$. S is an independent if not exists an edge between two vertices in S also exists in E. C is called vertex cover of G if and only if every edge uv ∈ E of G, at least one vertex, u or v in C. v in vertex in C, v is removable if with C - {v} is still a vertex cover of G. Denote $\rho(C)$ is the number of removable vertices of a C. a set of vertices called a Minimum vertex cover of a graph if has no removable vertices, mean that the least number of vertices.

Algorithm

[8] Description the algorithm with two procedures and two sub-procedures. Give graph G = (V, E), |V| = n; variable k (integer) is a threshold the number of vertices that cover all edges in a graph. If the vertex cover has size is k or small than k but cover all graph, algorithm stops. First sub-procedure find out the minimal vertex cover of C_i of each vertex in the graph. If C_i has no removable vertex, stop and output C. Else, with each removable vertex v of C, find $\rho(C-\{v\})$ of removable vertices of cover $C-\{v\}$. Denote v_{max} is the vertex such that $\rho(C-\{v_{max}\})$ is maximum. Then $C-\{v_{max}\}$ and repeat until there no removable vertices in the graph. In the second sub-procedure, with a minimal vertex cover C of G. If not exists a v in C connect to one neighbor $w \notin C$, output C. Else, find $v \in C$, which connect to w outside C and define $C^{v,w}$ by remove v from C then add w to C. Perform sub-procedure 1 with input is $C^{v,w}$. The result of sub-procedure 2 is resulting vertex cover.

Pseudocode procedure 1:

```
Input: G=(V,E)
Output: Minimal Vertex cover $C_i$
Begin
For i=1..n
 Initialize: $C_i = V - \{i\}$
 Sub_procecure_1 on $C_i$
 For r=1..n-k
        Sub_procecure_2
End;
```

Pseudocode procedure 2:

```
Input: Minimal Vertex cover $C_i$
Output: Minimal Vertex cover $C_{i,j}$
Begin
Initialize: $C_{i,j} = C_i \cup C_j$
Sub_procecure_1 on $C_{i,j}$
For r=1..n-k
 Sub_procecure_2
End;
```

Example

Give a G as **Figure *5*** (a). G have $C = V = \{1, 2, 3, 4, 5, 6, 7, 8\}$. Let caculate $C_1 = V - \{1\} = \{2, 3, 4, 5, 6, 7, 8\}$. Result number of removable vertices in **Table *1***, $\rho(C - \{v_{max}\})$ = 2 with v_{max}=3 and 6. Remove 3 out of C_1. Repeate the procedure with $C_1 = C_1 - \{3\} = \{2, 4, 5, 6, 7, 8\}$. $\rho(C - \{v_{max}\})$ = 0 with v_{max}=5 and 6. Remove 5 out of C_1 and stop. The result is minimum vertex cover with $C_1 = \{2, 4, 6, 7, 8\}$ with size = 5.

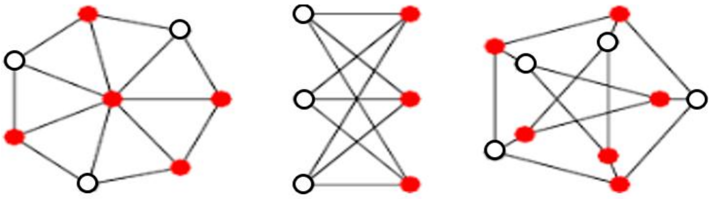
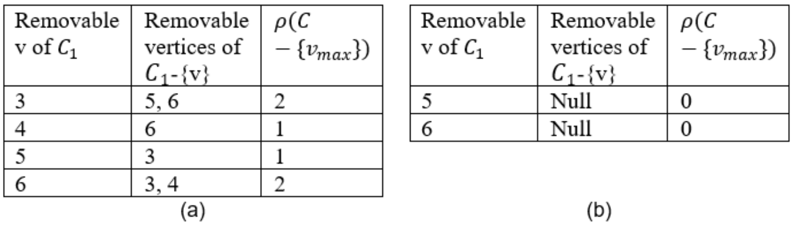


Figure 5. Graph find minimum vertex vertices

Table 1. C1 of graph Figure 4 (a)



Removable v of C_1	Removable vertices of C_1-{v}	$\rho(C - \{v_{max}\})$
3	5, 6	2
4	6	1
5	3	1
6	3, 4	2

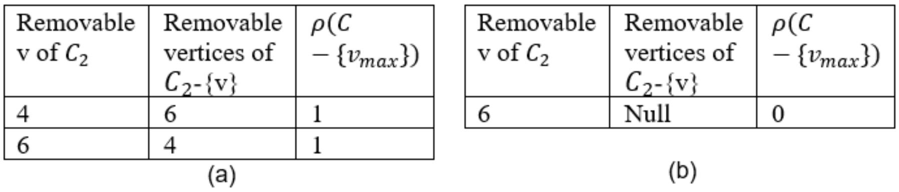
(a)

Removable v of C_1	Removable vertices of C_1-{v}	$\rho(C - \{v_{max}\})$
5	Null	0
6	Null	0

(b)

Give a G as **Figure *5*** (b). G have $C = V = \{1, 2, 3, 4, 5, 6\}$. Let caculate $C_2 = V - \{2\} = \{1, 3, 4, 5, 6\}$. Result number of removable vertices in **Table *2***, $\rho(C - \{v_{max}\})$ = 1 with v_{max}=4 and 6. Remove 4 out of C_2. Repeate the procedure with $C_2 = C_2 - \{4\} = \{1, 3, 5, 6\}$. $\rho(C - \{v_{max}\})$ = 0 with v_{max}= 6. Remove 6 out of C_2 and stop. The result is minimum vertex cover with $C_2 = \{1, 3, 5\}$ with size = 3

Table 2. C2 of graph Figure 4 (b)



Removable v of C_2	Removable vertices of C_2-{v}	$\rho(C - \{v_{max}\})$
4	6	1
6	4	1

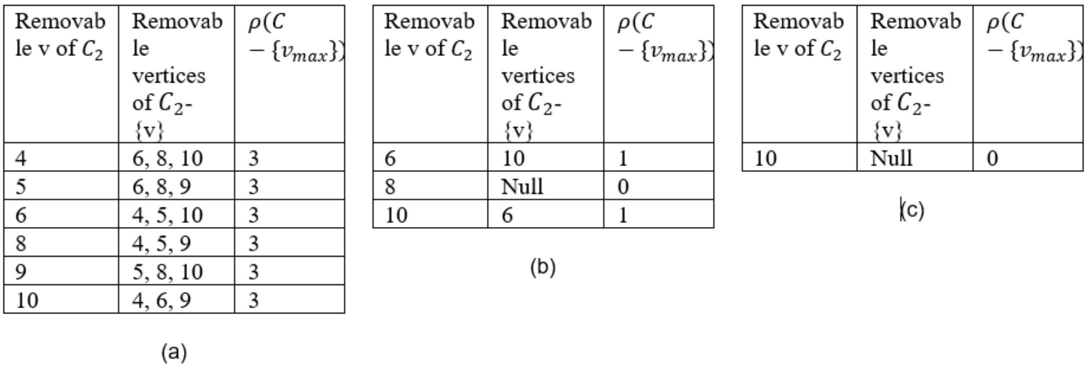
(a)

Removable v of C_2	Removable vertices of C_2-{v}	$\rho(C - \{v_{max}\})$
6	Null	0

(b)

Give a G as **Figure *5*** (c). G have $C = V = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. Let caculate $C_2 = V - \{2\} = \{1, 3, 4, 5, 6, 7, 8, 9, 10\}$. Result number of removable vertices in **Table *3***, $\rho(C - \{v_{max}\}) = 3$ with $v_{max} = 4, 5, 6, 7, 8, 9, 10$. Remove 4 out of C_2. Repeate the procedure with $C_2 = C_2 - \{4\} = \{1, 3, 5, 6, 7, 8, 9, 10\}$. $\rho(C - \{v_{max}\})$ = 1 with $v_{max} = 6, 10$. Remove 6 out of C_2 and stop. Continue with $C_2 = C_2 - \{6\} = \{1, 3, 5, 7, 8, 9, 10\}$. $\rho(C - \{v_{max}\}) = 0$. Remove 10 out of C_2 and stop. Minimum vertex cover with $C_2 = \{1, 3, 5, 7, 8, 9\}$ with size = 6.

Table 3. C2 of graph Figure 4 (c)



Removable v of C_2	Removable vertices of C_2-{v}	$\rho(C - \{v_{max}\}$
4	6, 8, 10	3
5	6, 8, 9	3
6	4, 5, 10	3
8	4, 5, 9	3
9	5, 8, 10	3
10	4, 6, 9	3

(a)

Removable v of C_2	Removable vertices of C_2-{v}	$\rho(C - \{v_{max}\}$
6	10	1
8	Null	0
10	6	1

(b)

Removable v of C_2	Removable vertices of C_2-{v}	$\rho(C - \{v_{max}\}$
10	Null	0

(c)

Determine k's length

There are much research metagenomic by using characteristics, probabilistic, Latent Dirichlet Allocation, almost use k-mer. As present in section 2, the size and content of words used to find features, the effect to the accuracy binning. As [2, 5], k = 4 is chosen, that means with four characters $\{A, G, T, C\}$ in DNA, there are only $4^4 = 256$ words or k-mer used to analyze. If k = 5, there is $4^5 = 1024$. To find the hidden model from sequences, which translate to document to apply LDA, the number of word in each document is $N_d = Length_d - k + 1$ ($Length_d = length\ of\ Sequence_d$). Example, with AGCTCTGAGA and k = 5, the document be translated AGCTC GCTCT CTCTG TCTGA CTGAG TGAGA. The size of a word is similar, the information in the sequence can mistake when split like this. The question is which k is suitable to build and how to make an effect to binning.

There are some problems need to consider when choosing k to build k-mer. First, if scan k from 1 to n (n's length < sequence's length), will generate a large of k-mers, include n number of 1-mer, 4^2 of 2-mer, 4^3 of 3-mer and go on. The number of k-mer very very big, Therefore, the time to look for features and binning data will increase dramatically. The idea to solve this problem is finding a minimum set of k-mer which have k's length different (but not full), still representation other k-mers. Second, mostly methods use k-mer, but not concert the order of k-mer, the new meaning when connecting directly from k-mers in sequence.

RESULT AND CONCLUSION

The algorithm in [8] used with metagenomic data in [2], species and sequences in **Table *4***. S$_i$ is short reads and R$_i$ is long reads. The experiments run with l-mer = 30 and threshold = 5 for both short and long reads. Others test with l-mer = 30 and threshold = 45 for long reads. The measures used in [2] continue used to test in this results, as bellow:

$$precicion = \frac{\sum_{i=1}^{k} max_j A_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{m} A_{ij}}$$

$$recall = \frac{\sum_{j=1}^{m} max_i A_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{m} A_{ij} + \#unassigned\ reads}$$

$$F - measure = \frac{2}{(\frac{1}{precision} + \frac{1}{recall})}$$

A_{ij}: number of reads from species j assigned to cluster i
K: number of cluster
M: number of species in metagenomics

Table 4. Specices and reads in experiment data

Data	Species	Ratio	Number of reads	Number of sequences
S1	2	1:1	96367	192734
S2	2	1:1	195339	390678
S3	2	1:1	338725	677450
S4	2	1:1	375302	750604
S5	3	1:1:1	325400	650800
S6	3	3:2:1	713388	1426776
S7	5	1:1:1:4:4	1653550	3307100
R1	2	1:1	82960	82960
R2	2	1:1	77293	77293
R3	2	1:1	93267	93267
R4	2	1:1	34457	34457
R5	2	1:1	40043	40043
R6	2	1:1	70550	70550
R7	3	1:1:8	290473	290473
R8	3	1:1:8	374830	374830
R9	6	1:1:1:1:2:14	588258	588258

Figure *6*, **Figure *7***, **Figure *8***, **Figure *9*** results of Recall, Precision and F-measre between BiMeta and MVC on short and long reads with l-mer = 30 and threshold (m) = 5.

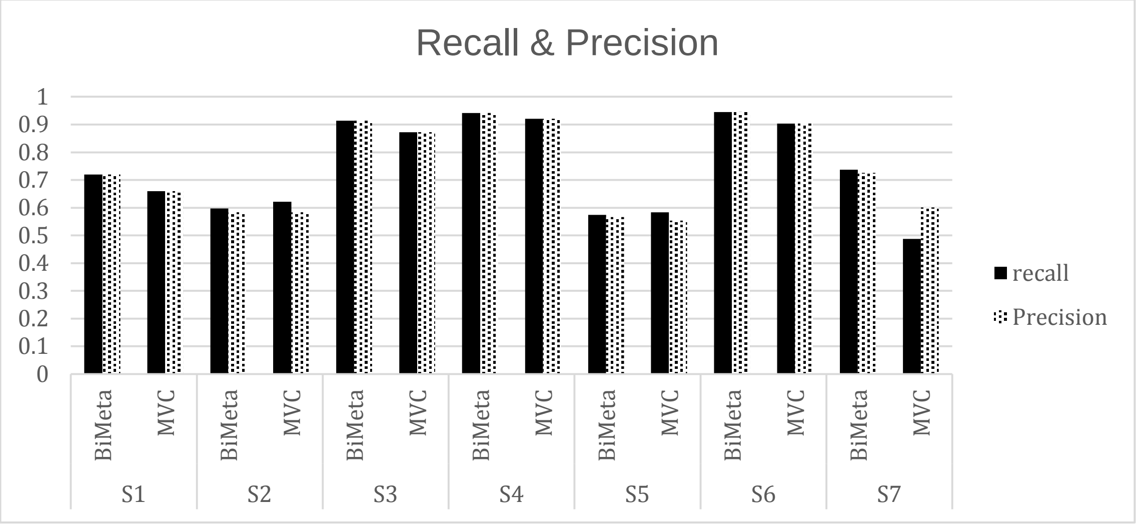


Figure 6. Recall & Precision short-reads (m = 5)

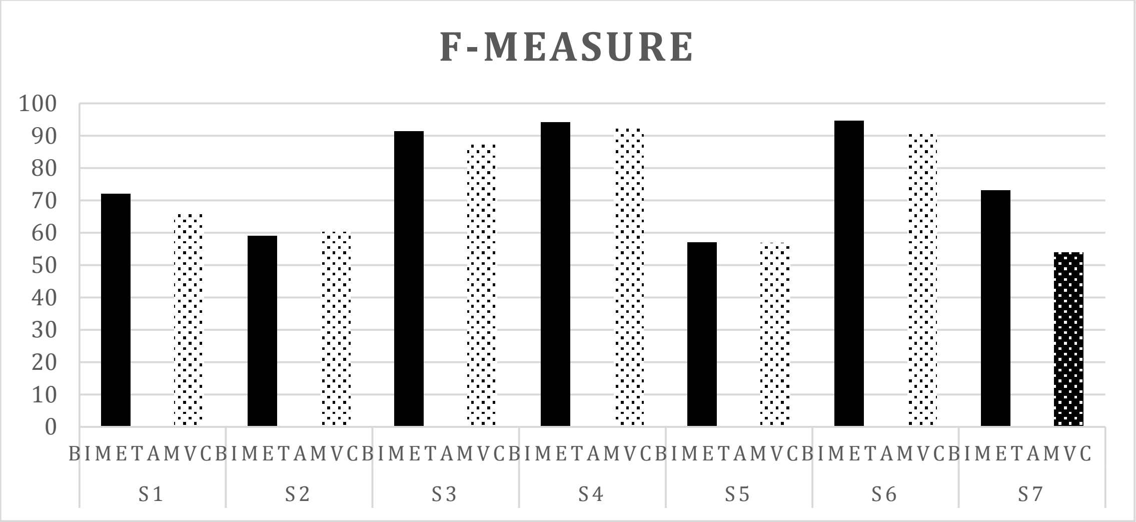


Figure 7. F-measure short-reads (m = 5)

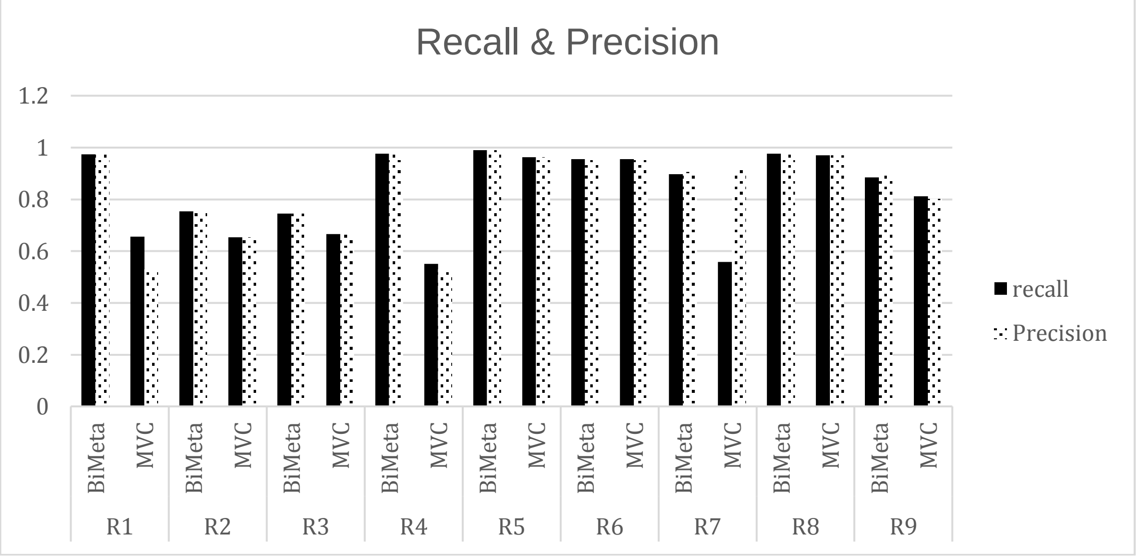


Figure 8. Recall & Precision long-reads (m = 5)

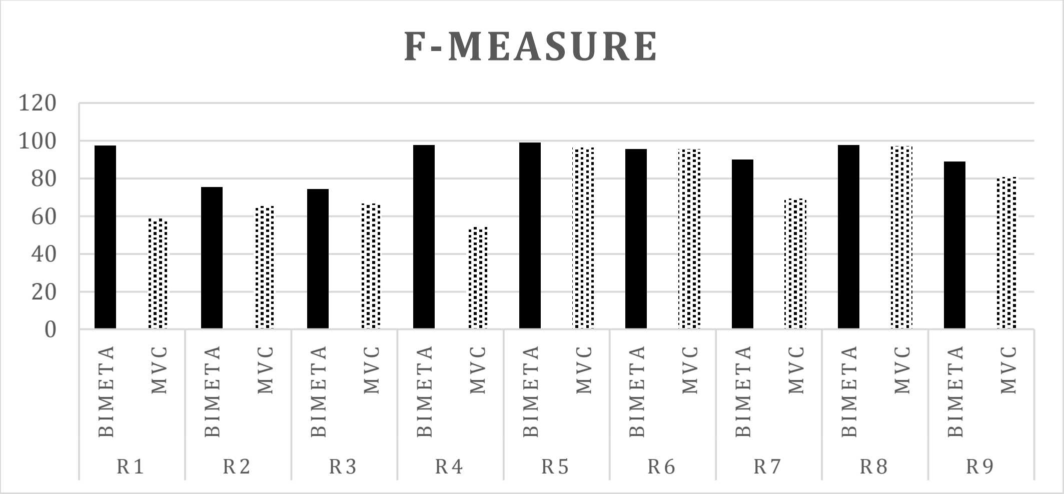


Figure 9. F-measure long-reads (m = 5)

Figure *10*, **Figure *11*** results of Recall, Precision and F-measre between BiMeta and MVC on long reads with l-mer = 30 and threshold = 45.

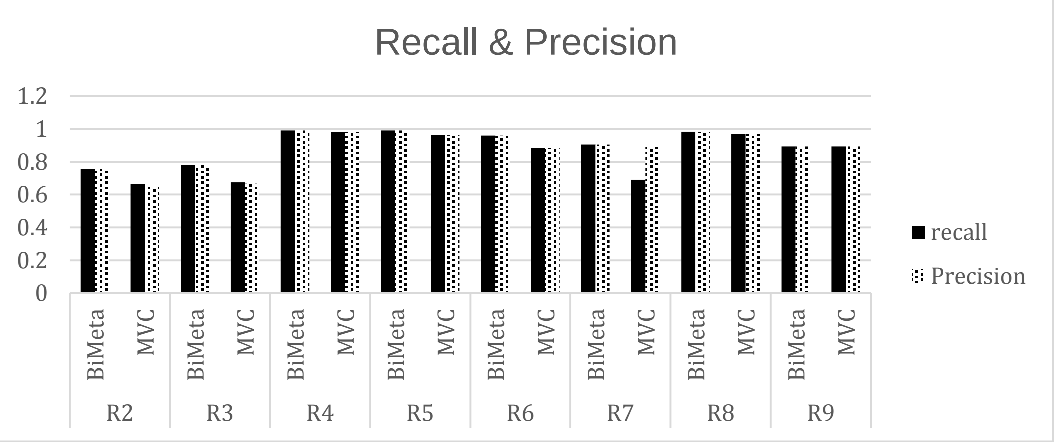


Figure 10. Recall & Precision long-reads (m = 45)

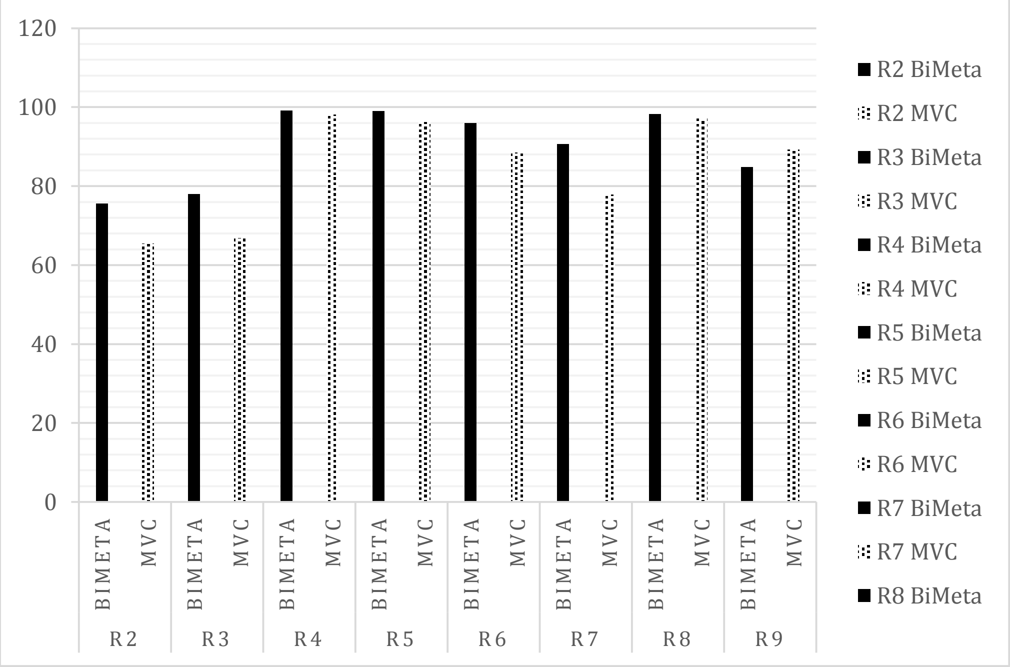


Figure 11. F-measure long-reads (m = 45)

The accelerated of studies in biological, specifically in microbial organisms raise the importance of Metagenomics. To find out and understand functions of various community genomics, still a lot of things to do, solve problems for each research direction. In this research, we review some technologies, characteristics of each methods using to cluster metagenomic data. Draw the problems exists and propose ideas to solve.

Find the suitable set of k-mer and minimum the data to find k-mer which used determine features to binning, promise to provide a greater additional to increate the accuracy and the performance. The idea expresses in section 3 is one of many methods used to mitigate existing limited in some of the stated methods. The algorithm is terminated in polynomial-time. It is importance to do an experiment when data of metagenomic increasing now. Another key of algorithm is conditions to always find the minimal vertex covers of graph.

REFERENCES

[1] R. Isea, "The present-Day Meaning of The Word Bioinformatics," *Global Journal of Advanced Research,* vol. 2, no. 1, pp. 70-73, 2015.

[2] Vinh LV, Lang TV, Binh LT, Hoai TV, "A two-phase binning algorithm using l-mer frequency on groups of non-overlapping reads," *Algorithms for Molecular Biology: AMB,* vol. 10, no. 2, 2015.

[3] Liu L, Tang L, Dong W, Yao S, Zhou W, "An overview of topic modeling and its current applications in bioinformatics," *SpringerPlus,* vol. 5, no. 1, p. 1608, 2016.

[4] Samuele Girotto, Cinzia Pizzi, Matteo Comin, "MetaProb: accurate metagenomic reads binning based on probabilistic sequence signatures," *Bioinformatics,* vol. 32, no. 17, 9 2016.

[5] Văn Đình Vỹ Phương, Trần Văn Lăng, Trần Văn Hoài, Lê Văn Vinh, "Áp dụng mô hình ẩn kết hợp thuật toán BiMeta trong việc gom nhóm trình tự metagenomic," in *Kỷ yếu Hội nghị Quốc gia về Nghiên cứu cơ bản và Ứng dụng CNTT lần thứ 9 - FAIR'9 Cần Thơ - ISBN: 978-604-913-472-2*, 2016.

[6] Zhang R, Cheng Z, Guan J, Zhou S, "Exploiting topic modeling to boost metagenomic reads binning," *BMC Bioinformatics,* vol. 16, no. (Suppl 5):S2, 2015.

[7] David M. Blei, Andrew Y. Ng, Michael I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research,* vol. 3, no. 4, pp. 993-1022, 2012.

[8] A. Dharwadker, The Vertex Cover Algorithm, Amazon, 2011.

VẤN ĐỀ GIẢM KÍCH THƯỚC TRONG VIỆC GOM NHÓM DỮ LIỆU METAGENOMIC

Văn Đình Vỹ Phương, Trần Văn Lăng, Trần Văn Hoài, Lê Văn Vinh

TÓM TẮT: *Bài báo xem xét các phương pháp gom nhóm dữ liệu metagenomic như: sử dụng k-mer để xác định các đặc trưng; sử dụng k-mer tìm mô hình ẩn trong dữ liệu. Sau đó, tiến hành gom nhóm trình tự dựa vào các đặc trưng hoặc mô hình tìm được. Để nâng cao hiệu suất, quá trình gom nhóm được thực hiện dựa trên các dữ liệu đại diện thay vì toàn bộ dữ liệu ban đầu. Việc tìm dữ liệu đại diện từ đó tìm tập đặc trưng cho toàn bộ dữ liệu gốc tiềm ẩn nhiều vấn đề. Bài báo trình bày ý tưởng sử dụng tập đỉnh bao phủ tối thiểu hoặc tập k-mer để đánh giá sự đúng đắn của dữ liệu đại diện.*

Từ khoá: *metagenomic, gom nhóm, trình tự, k-mer.*