IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 10, 2019, accepted November 25, 2019, date of publication December 5, 2019,
date of current version December 23, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2957833

VFM: Identification of Bacteriophages From Metagenomic Bins and Contigs Based on Features Related to Gene and Genome Composition

QIAOLIANG LIU[1], FU LIU[1], JIAXUE HE[2], MIAOLEI ZHOU[1], (Member, IEEE), TAO HOU[1], AND YUN LIU[1], (Member, IEEE)

[1]College of Communication Engineering, Jilin University, Changchun 130012, China
[2]Genetic Diagnosis Center, The First Hospital of Jilin University, Changchun 130021, China

Corresponding author: Yun Liu (liuyun313@jlu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61503151, in part by the Natural Science Foundation of Jilin Province under Grant 20160520100JH, and in part by the China Postdoctoral Science Foundation under Grant 2019M651204.

ABSTRACT As the main regulator of microbial community composition, bacteriophages exist widely on Earth. However, since they are hidden in metagenomes, most of them are unknown. To identify phages from metagenomes more effectively, a new tool named VFM (Virus Finding & Mining) is presented in this paper. VFM has two versions, i.e., bin-VFM and unbin-VFM. Eighteen new features describing the codon usage bias, the proportion of hits of clusters of orthologous groups of proteins (COG), and 1-mer and 2-mer frequency are introduced to improve the performance of the classifiers. By using missing value interpolation, bin-VFM improves the classification performance for short sequence bins significantly. Compared with previous tools for virus mining, bin-VFM and unbin-VFM perform much better for simulated and real metagenomes with short and long sequences respectively. Thus, VFM may play a helpful role in studies of metagenome-related problems, such as horizontal gene transfer and antibiotic resistance. VFM is freely available at https://github.com/liuql2019/VFM.

INDEX TERMS Codon usage bias, COG, missing value interpolation, metagenomic viruses, phage mining, short k-mer frequency.

I. INTRODUCTION

Viruses are the most abundant and widespread life forms on Earth [1]. Their habitats include host bodies [2], such as humans [3]–[5], animals [6]–[8], insects [9], and plants [10], as well as natural environments [11], including marine [12]–[14], freshwater [15], springs [16], soil [17]–[19], and other niches [20], [21]. In the metagenomes [22] obtained from these habitats, the majority of viruses are bacteriophages, which have great impacts on the composition and function of the microbial floras and then affect the host bodies and surrounding environments [23]–[25]. Antibiotic resistance genes have been proved to be associated with bacteriophages infecting the microbes of humans [26]. Horizontal gene transfer among microbial species is also regulated by their parasitic phages [27]. Therefore, the identification of phage sequences from a variety of metagenomes plays a crucial role in metagenomic research.

The associate editor coordinating the review of this manuscript and approving it for publication was Quan Zou.

However, compared with prokaryotic and eukaryotic organisms, information on relatively few viruses has been deposited in the current biological databases. Since viruses have no universal marker genes that characterize them, it remains difficult to discover novel viruses hidden in metagenomes. The unknown sequences in metagenomes, which include a large number of novel viruses, are aptly described as "dark matter" [28]. Tools such as VIP [29], VirusSeeker [30], ViromeScan [31], and FastViromeExplorer [32] have been designed for the identification of virus reads which are based on homology comparison with known organisms in databases; these tools perform well for mapping virus reads to known viruses, whereas novel viruses are hardly discovered. Applications such as virSorter [33], virMine [34], virFinder [35], and virMiner [36] focus on contig identification of phages in metagenomes. virSorter makes use of subjective guidelines to make decisions, while the subtleties of features are difficult to detect. Using machine learning, virFinder uncovers viruses based on 8-mer frequency [37]; virMiner uses features calculated by gene hits to several gene databases, which exploits contigs from actual

human metagenomes as training data that may result in species bias and mislabeling of the training data. By relying on gene information, MARVEL [38] benefits from the binning technique [39], [40] to ensure better classification performance for the bins of long contigs, where its performance for bins consisting of short contigs requires improvement. Furthermore, binning process may lead to some hybrid incorrect bins and the single contig from some low abundance species cannot be binned in some cases.

Here we propose a machine learning-based detector named VFM with two versions. In addition to six features used by MARVEL, another eighteen new features are used, including five features related to codon usage bias, one feature related to the proportion of COG gene hits, two features related to 1-mer frequency, and ten features related to 2-mer frequency. The bin version named bin-VFM aims to find phages more accurately from metagenomic bins, especially for bins consisting of short contigs. To achieve this goal, for bins with only one gene at most, the mean values of the features related to two or more genes in the training set are inserted to deal with the missing value problem. The other version (unbin-VFM) uses the same features as bin-VFM to handle unbinned sequences, especially long contigs. Both versions provide state-of-the-art performance compared with other tools previously designed for phage mining.

II. MATERIALS AND METHODS

A. SIMULATED BINS AS TRAINING AND TEST SETS

The dsDNA phage and bacterial genomes were obtained from the NCBI RefSeq database. The species in the training sets consisted of 1247 phages in the Caudovirales order and 1029 bacteria released before January 1, 2016; 8 kbp sequences were sampled randomly from those genomes to generate simulated bins, each of which contained 10 sequences derived from the same genome [38]. To simulate bins most of which had no genes, another training set was created. Each bin contained 10 sequences with the length of 500 bp. In order to cover more regions of a genome, 500 bp sequences were selected 10 times randomly from the same genome to generate 10 bins.

The test sets contained phages in the Caudovirales order and bacteria released after January 1, 2016; 200 phages and 400 bacteria were chosen at random from these genomes for the test sets. Seven test sets were created by using sequences with seven different lengths (0.5 kbp, 1 kbp, 2 kbp, 3 kbp, 4 kbp, 8 kbp, and 12 kbp). For each length, 10 sequences were selected randomly from every chosen genome to simulate a bin in the test set. The training and test sets are available at https://www.jianguoyun.com/p/DYIe6QgQ7I_kBxihkPUB.

B. EXTRACTING FEATURES FROM SIMULATED BINS

24 features were chosen to create a feature vector for every simulated bin. These features were divided into three categories based on gene statistics, gene coding, and oligonucleotide usage frequency respectively. The features in the first category included gene length, strand shift frequency [41], spacing size, gene density, and ATG frequency [38]. The features in the second category included five features related to codon usage frequency bias [42], two features describing the proportion of pVOG [43] gene hits, and the proportion of COG [44] gene hits among all genes. One-fourth of the COG gene database was selected at random after removing some genes of viral origin [34]. The features in the third category included 1-mer usage frequency and 2-mer usage frequency for the overall coding and non-coding regions. The former consists of two-dimensional features, whereas the latter is comprised of ten-dimensional features. In detail, 1-mer frequency counted the frequency of A and G in a sequence with T and C omitted as the number of T is equal to A and the number of C is equal to G. For the same reason, 2-mer frequency counted the frequency of AA, AT, AG, AC, TA, TC, TG, CG, CC, GC. Each feature of a bin was calculated based on the weighted average of the corresponding feature from all contigs in the bin.

Considering that different sequence lengths exist in a real metagenomic bin, weight parameters were used for the feature average calculation. The weights corresponding to the above mentioned features were as follows. The weight for the gene length was the number of genes in the sequence. The weight for spacing size was the number of spaces between every two genes. The weight for gene density was the sequence length calculated by base. The weight for strand shift frequency was the number of spaces between two genes. The weight of the ATG frequency was the sequence length minus two. All five weights of the features indicating codon usage frequency bias were the sum of all gene lengths in the sequence. All weights for the 1-mer frequency features were the same, i.e., the sequence length, whereas, for the 2-mer frequency features, the weights were the sequence length minus one. In order to address the problem that many 0.5 kbp sequences have no genes, 4-mer frequency was chosen as the features of the 0.5 kbp training set. The 4-mer frequency was calculated for every sequence in a bin prior to calculating their weighted average. The sequence length minus three was used as the weight parameters.

C. TRAINING AND TESTING FOR SIMULATED BINS

The classifier trained for phage bin prediction is bin-VFM. Following feature acquisition from the 0.5 kbp and 8 kbp training sets, two models were trained using these features. Logistic regression and random forest algorithms in Python3.5 were used for the 0.5 kbp and 8 kbp training sets respectively. For the logistic regression, the value of the C parameter, which determines the L2 regularization degree, was 10^{-7}, which was acquired by the trial on a validation set created using a small portion (slightly more than 1/10) of bins in the training set. Number 50 was chosen as the tree number for random forest algorithm.

In the initial stages of the prediction process, one of the two models should be chosen based on whether the bin being predicted contains gene(s). The logistic regression model was

applied to the bins without genes using 4-mer frequency features. In contrast, the random forest model was applied to the bins with one gene at least in it using the 24 features. For short contig bins, some features related to at least two genes, strand shift frequency and spacing size in detail, would be missing values. However, feature vectors with missing values can't be handled by the model. Prior to the prediction, such null values were filled with missing value interpolation based on mean values of the features of the training data. Specifically, the null values of the two features were filled by the mean value of the corresponding feature in the training set, respectively, which improved the performance for short contig bins. The prediction results showed whether the bins were phage bins or not, followed by the probabilities as a phage bin. Fig.1 is the program flow chart of bin-VFM.

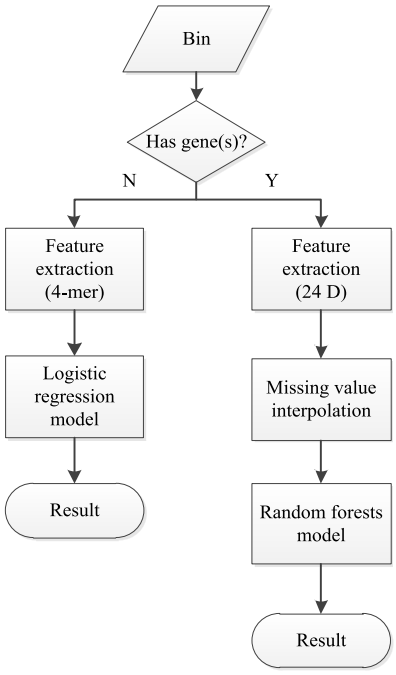

FIGURE 1. Program flowchart of bin-VFM.

D. TRAINING AND TEST SETS FOR SIMULATED UNBINNED CONTIGS

The genomes of phages (Caudovirales order) and bacteria released before January 1, 2016 were split based on three different lengths (2 kbp, 4 kbp, and 8 kbp) to create three unbinned training sets. To maintain a balance between phages and bacteria, the bacterial sequences were selected randomly to be equal to the phage sequences. The genomes used for the test sets of Bin-VFM were split into sequences with lengths of 2 kbp, 3 kbp, 4 kbp, 8 kbp, and 12 kbp for five unbinned test sets. Since the number of bacteria exceeds phages in real metagenomes [45], the test sets were created by selecting 2000 phage sequences and 4000 bacterial sequences for each certain length. The unbinned training and test sets are also available at https://www.jianguoyun.com/p/DYIe6QgQ7I_kBxihkPUB.

E. TRAINING AND TESTING FOR SIMULATED UNBINNED CONTIGS

The classifier trained for the phage contig prediction is named unbin-VFM. Corresponding to the lengths of 2 kbp, 4 kbp, and 8 kbp, three models were trained using the random forest algorithm and the same 24 features as for bin-VFM. To utilize the training program developed for bin-VFM, one file was created for each contig as a single contig bin. The simple bins of the three training sets were dealt with as the training process of bin-VFM to generate three models for the lengths of 2 kbp, 4 kbp, and 8 kbp respectively. For the shorter contigs in the training and test sets, some features related to genes may be missing values. Missing value interpolation was executed and the averages of the corresponding feature values in the training set were used to fill in the missing values. The 2 kbp model was developed for predicting contigs shorter than 4 kbp, the 4 kbp model was used for contigs between 4 kbp and 8 kbp, and the 8 kbp model was used for contigs longer than 8 kbp. All contigs in the unbinned test sets were predicted by the three models based on the lengths of the contigs.

F. COMPARISON WITH OTHER TOOLS USING THE SIMULATED TEST SETS

In order to evaluate the performance of VFM for the test sets with different lengths of bins or contigs, other tools described in previous publications were used for comparison, i.e., virSorter, virFinder, and MARVEL. The same test sets were used for all tools to conduct a fair comparison and evaluate the ability of the tools for phage detection. The sequences with a certain length may be inappropriate for some tools. MARVEL cannot handle 0.5 kbp bins without genes. In this case, MARVEL would not participate in the evaluation for the 0.5 kbp test set.

Five metrics, namely, recall, specificity, precision, accuracy, and the F1 score were calculated to quantify the performance of the tools. The equations for calculating the metrics are described as follows.

$$Recall = TP/P \tag{1}$$

$$Specificity = TN/N \tag{2}$$

$$Accuracy = (TP + TN)/(P + N) \tag{3}$$

$$Precision = TP/(TP + NP) \tag{4}$$

$$F1 = 2 \times (Precision \times Recall)/(Precision + Recall) \tag{5}$$

In these equations, *P, N, TP, FP, TN, FN* are the number of real positive cases, the number of real negative cases, the number of true positives, the number of false positives, the number of true negatives, and the number of false negatives respectively in final results, where phages are labeled as positive and bacteria as negative. F1 is a better index than accuracy for imbalanced test sets.

G. PERFORMANCE EVALUATION USING REAL METAGENOMES

Some differences exist between simulated contigs and real metagenomic contigs; therefore, the performance of VFM was investigated using real metagenomes. Contigs created by cross-assembly of the Amazon River plume samples from the Amazon dataset [46] were used for testing. The assembly tool was Rey 2.3.1 [47] using 31-mer. Bin-VFM and MARVEL were chosen for the test of bin prediction. Before binning, contigs shorter than 500 bp were removed. COCACOLA [48] was used for implementing the binning tasks in which the cluster number was set to 500. BLASTn [49] was used to assign class labels to the contigs in these bins as a benchmark by alignment with the genomes of phages and bacteria released before December 1, 2018 in the NCBI Refseq database. For the unbinned contig test of unbin-VFM and virFinder, 1% of the metagenomic viral contigs (mVCs) [1] were selected randomly for recall test and the environmental contigs (>= 2kbp) that were labeled as bacteria previously for the bin prediction test were chosen for specificity test.

III. RESULTS

A. THE OVERALL DESIGN OF VFM

To achieve better performance for phage mining, eighteen new features related to codon usage bias, COG gene ratio, and short k-mer frequency (k = 1,2) were used to create a new longer feature vector based on the six features previously reported [38], [41] (see Method). Codon usage bias refers to the fact that different species often have distinct synonymous codons in their genes; this can be used as a gene-related marker for identification [42]. The clusters of orthologous groups of proteins (COG) database consists of a large number of microbial orthologous genes; it has been updated since 2014 and can be used to identify prokaryotic sequences [44]. Considering that the large volume of the COG database would slow down processing speed, a smaller COG database consisting of 1/4 of the original genes (some virus shared genes were deleted [34]) is used for the COG gene ratio calculation. The reason for using 1-mer and 2-mer frequency is that the short k-mer frequency can also facilitate phage mining tasks.

With the random forests algorithm, one model was trained using 8 kbp bins for bin-VFM and three models were trained using contigs with lengths of 2 kbp, 4 kbp, and 8 kbp for unbin-VFM. All the bins and contigs for training originated from ssDNA phages and bacteria released prior to January, 1, 2016; the data were downloaded from the National Center for Biotechnology Information (NCBI). When bins or contigs are predicted, for the bins with contigs or the unbinned contigs that are long enough to contain two or more genes in one sequence, all 24 features can be calculated. However, some features related to two or more genes should be supplemented for the bins with one-gene contig(s), which are sometimes mixed with no-gene contig(s), as well as unbinned short contigs with one or no genes. In order to solve this problem, the missing feature(s) of the bins or contigs are preprocessed by missing value interpolation and the averages of the corresponding values in the training set are used to fill in the missing values. In particular, for bins consisting of only no-gene contigs, 4-mer frequency features are used instead of the 24 features because most of the 24 features are incalculable in such cases. The no-gene bins are predicted by a specially trained classifier using the features of 4-mer frequency. In addition, all the feature vectors of the contigs in a bin are fused by calculating the weighted means to weigh their contributions to the bin. After feature extraction, all feature vectors are used by either the bin-VFM model or corresponding unbin-VFM model for phage prediction.

B. COMPARISON WITH OTHER TOOLS USING SIMULATED BINS AND CONTIGS

Bin-VFM and unbin-VFM are the two versions of VFM. Bin-VFM was developed for bins consisting of contigs with various lengths, whereas unbin-VFM was for single unbinned contigs. To evaluate VFM for classifying bins or contigs with different lengths, bin-VFM was compared with MARVEL and virFinder using simulated bins with sequence lengths ranging from 0.5 kbp to 12 kbp; unbin-VFM was compared with virFinder and virSorter using simulated contigs with lengths ranging from 2 kbp to 12 kbp. All the sequences of the simulated bins and the unbinned contigs were sampled randomly from 200 phage and 400 bacterium genomes released after January 1, 2016. Recall, specificity, precision, accuracy, and the F1 score were used to evaluate their performance. VFM demonstrated better performance than the other tools for both bins and contigs for nearly all lengths.

Bin-VFM, MARVEL, and virFinder were evaluated using simulated bins with seven contig lengths. Since MARVEL predicts phage bins with the probability threshold of 70%, for a fair comparison, the same threshold was chosen for bin-VFM and virFinder. Bin-VFM nearly outperformed both MARVEL and virFinder for each evaluation score (Fig.2). In particular, for recall, precision, accuracy, and F1, bin-VFM performed much better than the other tools on the shorter lengths (<4 kbp). In the best case, for the 1 kbp length, bin-VFM performed ∼40%, ∼20%, and ∼30% better than MARVEL for recall, accuracy, and F1 respectively. For sequences longer than 4 kbp, bin-VFM exhibited nearly the same predictive ability as MARVEL and both performed better than virFinder. Additionally, MARVEL failed to handle 0.5 kbp bins since there was no gene in some of the bins. For specificity, bin-VFM and virFinder achieved similar values, whereas MARVEL performed slightly better for the length of 1 kbp but had much lower recall than bin-VFM for the same length.

Unbin-VFM, virFinder, and virSorter were run on simulated contigs derived from the phages and bacteria released after January 2016. Unbin-VFM achieved the best results for all evaluation criteria (Fig.3). Similar to previous studies, virSorter did not perform well for each of the criteria.

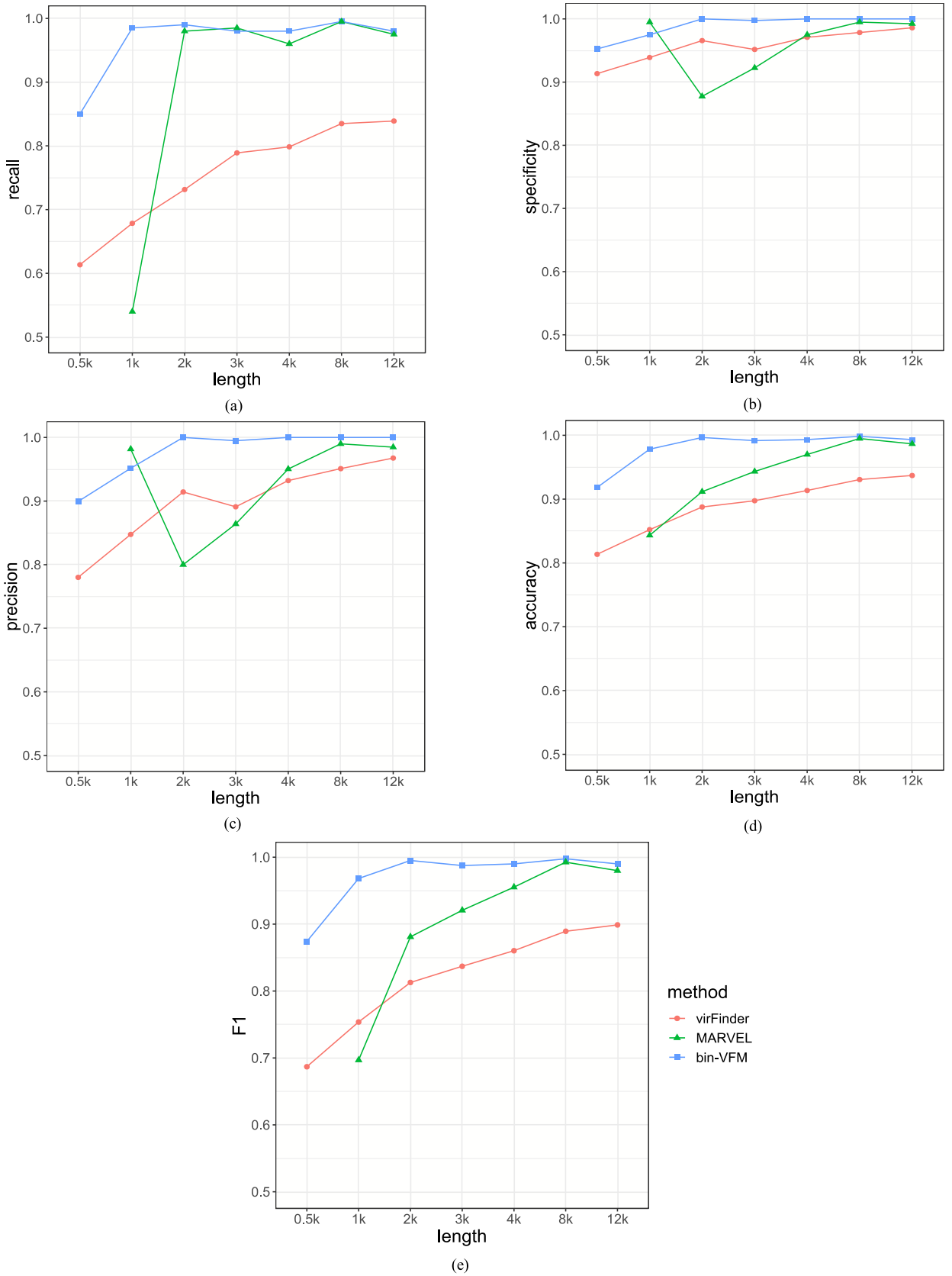

FIGURE 2. Prediction results of bin-VFM, MARVEL, and virFinder for simulated bins.

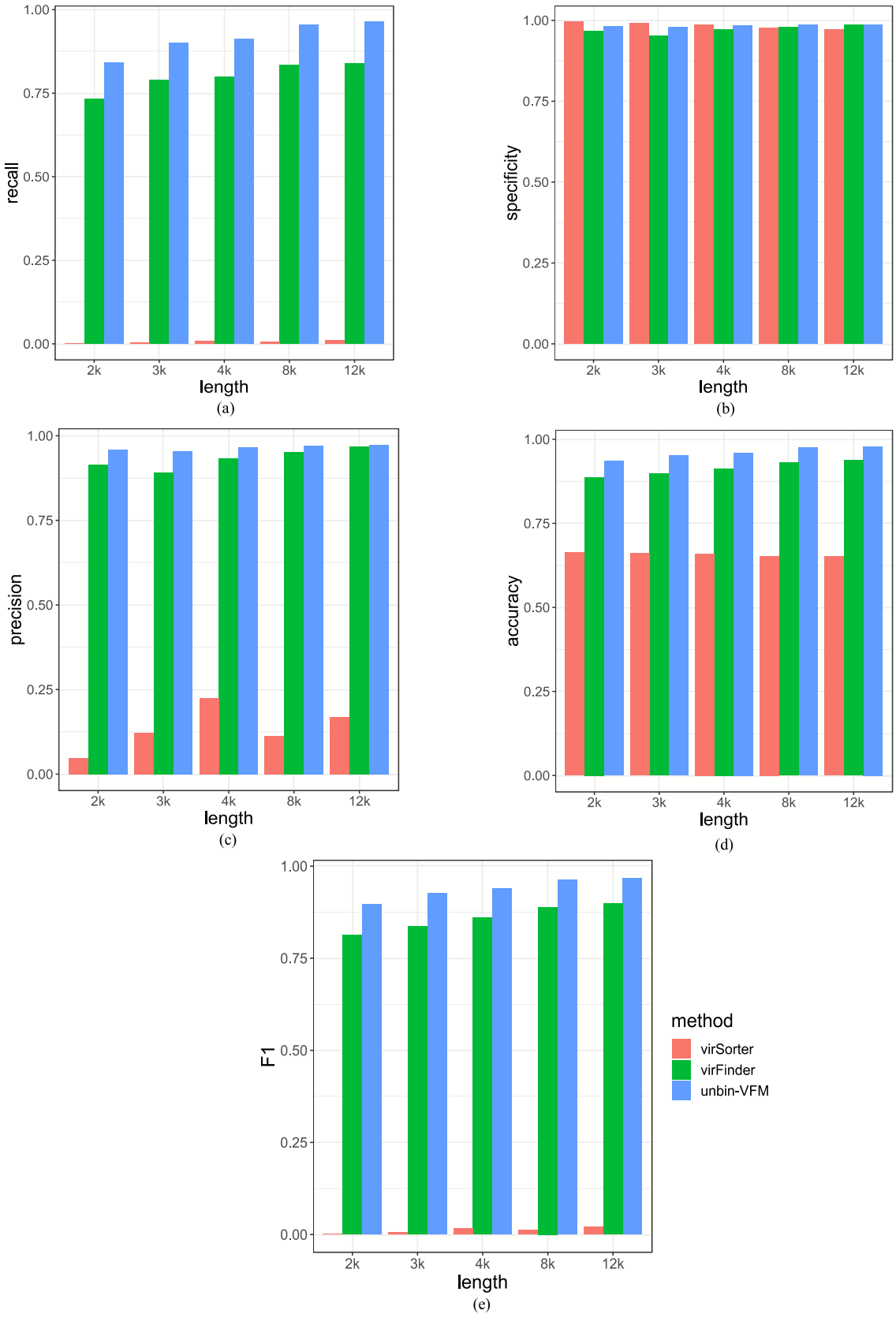

FIGURE 3. Prediction results of unbin-VFM, virFinder, and virSorter for simulated contigs.

The recall score of unbin-VFM was ∼10% higher than virFinder. For accuracy and F1, the scores of unbin-VFM were ∼5% and ∼8% higher than virFinder, indicating better comprehensive capability for phage prediction. For the precision score, unbin-VFM achieved better performance for short contigs than virFinder. Moreover, the shorter the contigs, the better the result was. Similar to the simulated bin results, nearly perfect specificity was achieved by all tools. It was remarkable that unbin-VFM had higher specificity scores than virFinder for all lengths, indicating that unbin-VFM will provide higher precision for real metagenomes because the proportion of bacteria may be much larger than phages in real metagenomes.

C. PREDICTION RESULTS OF TOOLS ON REAL METAGENOME

In order to evaluate the performance of VFM to process real metagenomes, the samples from a natural environment were used [46]. After being processed via cross-assembly, high-quality contigs were obtained. After binning the contigs, bin-VFM and MARVEL were compared to evaluate their performance. Given that it is very difficult to acquire correct class labels of contigs in real metagenomes, a reliable method was chosen to determine the labels by using BLASTn to align the contigs to the genomes of phages and bacteria released before December 1, 2018. Since MARVEL cannot process bins without genes, 5 bins without genes were not used in the MARVEL test set. The results are shown in Table 1. We can conclude that bin-VFM performed better than MARVEL for this real metagenome. Bin-VFM achieved ∼10% higher recall than MARVEL. Because bacteria are more abundant than phages in real metagenomes, the 2% higher specificity of bin-VFM indicates that a large number of bacteria have been filtered out. Therefore, the precision showed the maximum percentage difference of all evaluation criterions between bin-VFM and MARVEL at ∼13%.

TABLE 1. Prediction results of bin-VFM and MARVEL for real metagenome.

Tool	Recall	Specificity	Precision	Accuracy	F1
Bin-VFM	84.24%	97.14%	73.18%	96.05%	78.33%
MARVEL	74.60%	95.56%	60.89%	93.78%	67.05%

A large number of metagenomic viral contigs (mVCs) have been discovered from thousands of real metagenomes by Paez-Espino *et al.* [1]; these can be used to evaluate phage mining tools. Since the number of contigs is too large, 1% of them was selected by random sampling to test the recalls of unbin-VFM and virFinder. The result showed that the recall of unbin-VFM (73.95%) was much better than virFinder (67.83%) for the same probability threshold (70%). Given the wide distribution of mVCs, we conclude that unbin-VFM has a better ability of recalling viruses than virFinder generally. In order to evaluate the effect of filtering out bacteria, BLASTn was used to label the contigs longer than 2 kbp. The contigs labeled as bacteria were used for assessing specificity. For the real metagenome, unbin-VFM had a specificity of 98.35%, where the score of virFinder was slightly lower at 97.80%. This indicated that for real contigs, both tools could perform well in filtering out bacteria.

IV. DISCUSSION

Here we propose two classifiers named bin-VFM and unbin-VFM for binned and unbinned contigs to identify dsDNA phages from metagenomes. Eighteen new features were introduced and both bin-VFM and unbin-VFM outperformed their competitors. Moreover, since bin-VFM takes advantage of missing value interpolation and short k-mer frequency features, which ignore genes, it may better than MARVEL that was also designed for metagenomic bin prediction, especially for bins consisting of short contigs. When performing binning for real metagenomes, some short contig bins may be generated and bin-VFM is more useful than MARVEL in this case. That is one of the reasons why bin-VFM performed better than MARVEL for real metagenomes. In addition, since bins may contain contigs of different lengths in real metagenomes, the weighted average of contig feature vectors for a bin is used. That is another reason why bin-VFM had better performance than MARVEL. In unbinning cases, unbin-VFM is the first tool using the features of gene statistics, gene coding, and oligonucleotide usage frequency for phage contig prediction. Therefore, unbin-VFM showed better performance than the other tools, except for short contigs. For both simulated contigs and real contigs in the metagenome for test, the prediction results of unbin-VFM were clearly better than the comparative tools, indicating that unbin-VFM can detect more phage contigs.

TABLE 2. Recall of virFinder and unbin-VFM for three new randomly selected mVCs datasets.

Tools	1% of mVCs	0.7% of mVCs	0.5% of mVCs	Average
virFinder	64.81%	68.22%	70.63%	67.89%
unbin-VFM	73.31%	73.21%	73.02%	73.18%

To further investigate the recall for unbin-VFM, contigs were selected randomly (1%, 0.7%, and 0.5% of mVCs) to create three other virus sets for test. The result showed (Table 2) that unbin-VFM outperformed virFinder in all sets. On average, the recall of unbin-VFM was ∼6% higher than that of virFinder. We chose two additional probability thresholds (50% and 60%) to check the recall and specificity for unbin-VFM and virFinder (Fig.4, Fig.5). The results were consistent with the expectation. Unbin-VFM outperformed virFinder in all cases and the smaller the threshold, the higher the recall and the lower the specificity were. For shorter contigs, the test of longer contigs (16 kbp and 24 kbp, Table 3) showed that unbin-VFM still performed much better.

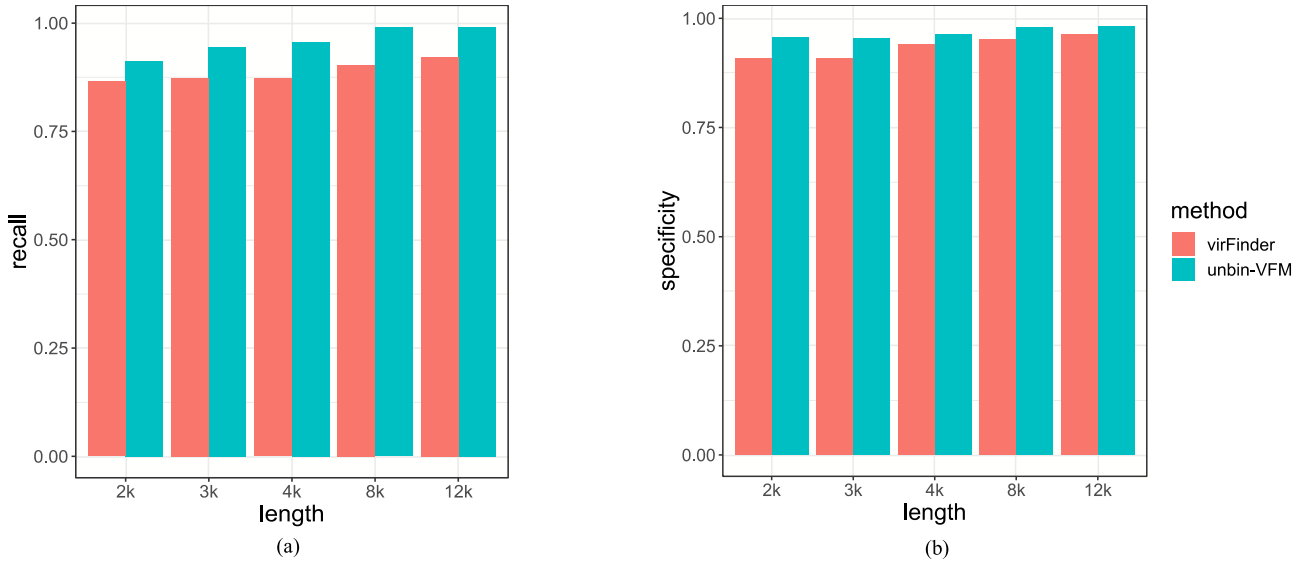

FIGURE 4. Performance of virFinder and unbin-VFM for simulated contigs at the probability threshold 50%.

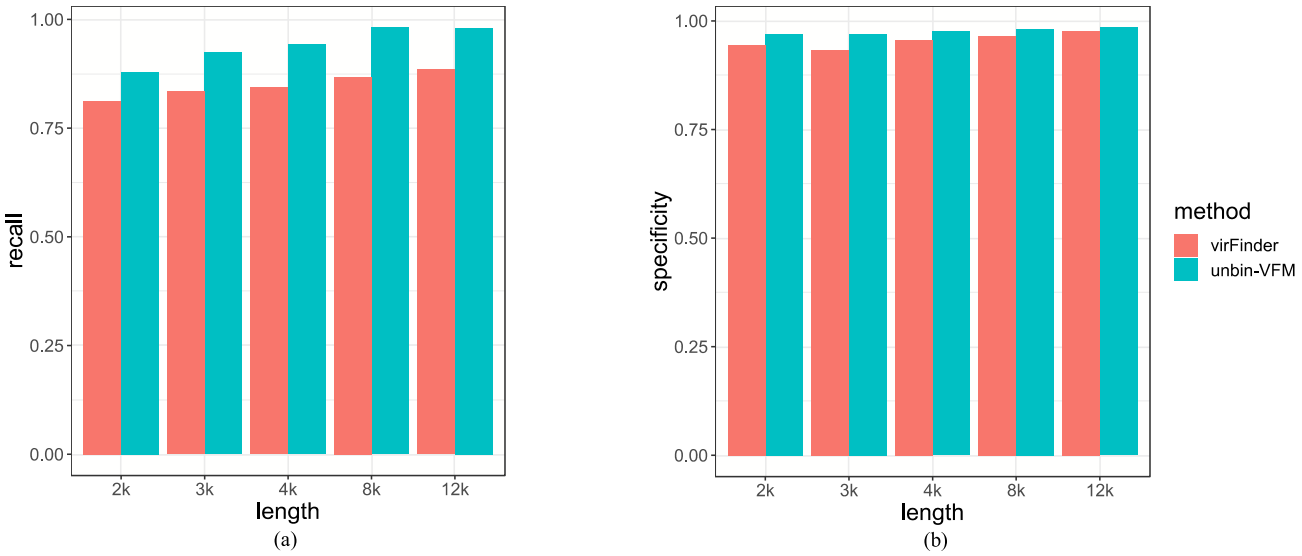

FIGURE 5. Performance of virFinder and unbin-VFM for simulated contigs at the probability threshold 60%.

VFM was designed only for predicting tailed phages belonging to the Caudovirales order, which is also a limitation of MARVEL. However, since viral metagenome communities mainly consist of tailed phages [50], the phage data for training is appropriate for the task of metagenomic virus mining. Furthermore, the recall test for mVCs, which contains a variety of dsDNA and ssDNA viruses from real metagenomes, confirms the predictive ability of unbin-VFM.

In the end, there are some limitations in the two versions of VFM. First, they are designed based on contigs, which may include chimeric sequences. In addition, binning processes may also generate a few erroneous bins containing contigs originating from different species. However, in order to discover novel and more complete viruses, it is necessary to use assembly and binning. Second, because most features of unbin-VFM are related to the computer-detected gene(s) in contigs, the performance of unbin-VFM may be worse for short contigs, such as contigs shorter than 1 kbp. We also tested its performance for 1kbp simulated contigs. Overall, the performance of unbin-VFM is comparable to that of virFinder (Table 3). Although its recall is slightly lower, the higher specificity compensates for that shortcoming. Third, the pVOG [43] database released in 2017 was created based on most of known viruses, which may reduce the effect of novel virus prediction. We believe that the variation of the difference of recall scores between unbin-VFM and virFinder for the simulated and real metagenomes may be attributed to this reason. It is hoped that researchers will be able to solve these problems in the future.

V. CONCLUSION

In summary, the two versions of VFM developed for classification of dsDNA phages and bacteria showed better performance than all similar tools in their respective fields.

TABLE 3. Performance of virFinder and unbin-VFM for shorter and longer simulated contigs.

Tools	length	Recall	Specificity	Accuracy	Precision	F1
virFinder	1kbp	67.85%	93.9%	85.22%	84.76%	75.37%
unbin-VFM	1kbp	63.70%	97.25%	86.07%	92.05%	75.30%
virFinder	16kbp	82%	98.4%	92.93%	96.24%	88.55%
unbin-VFM	16kbp	97.15%	98.65%	98.15%	97.30%	97.22%
virFinder	24kbp	82.9%	98.78%	93.48%	97.13%	89.45%
unbin-VFM	24kbp	96.05%	98.65%	97.78%	97.27%	96.65%

The proposed tools may be helpful for researchers to uncover more novel viruses in the form of contigs and bins from rapidly growing metagenomes. Previously developed tools can also be used in conjunction with bin-VFM and unbin-VFM to exploit their own advantages. We believe that the use of these tools by virology researchers may result in the discovery of a great number of novel viral taxa and would further have a marked impact on virus-related research, such as human health and environmental problems.

REFERENCES

[1] D. Paez-Espino, E. A. Eloe-Fadrosh, G. A. Pavlopoulos, A. D. Thomas, M. Huntemann, N. Mikhailova, E. Rubin, N. N. Ivanova, and N. C. Kyrpides, "Uncovering Earth's virome," *Nature*, vol. 536, no. 7617, pp. 425–430, Aug. 2016.

[2] K. Cadwell, "The virome in host health and disease," *Immunity*, vol. 42, no. 5, pp. 805–813, May 2015.

[3] L. A. Ogilvie and B. V. Jones, "The human gut virome: A multifaceted majority," *Frontiers Microbiol.*, vol. 6, p. 918, Sep. 2015.

[4] B. E. Dutilh, N. Cassman, K. McNair, S. E. Sanchez, G. G. Z. Silva, L. Boling, J. J. Barr, D. R. Speth, V. Seguritan, R. K. Aziz, B. Felts, E. A. Dinsdale, J. L. Mokili, and R. A. Edwards, "A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes," *Nature Commun.*, vol. 5, p. 4498, Jul. 2014.

[5] J. Wang, Y. Gao, and F. Zhao, "Phage-bacteria interaction network in human oral microbiome," *Environ. Microbiol.*, vol. 18, no. 7, pp. 2143–2158, Jul. 2016.

[6] D. J. Obbard, "Expansion of the metazoan virosphere: Progress, pitfalls, and prospects," *Current Opinion Virol.*, vol. 31, pp. 17–23, Aug. 2018.

[7] M. Carda-Dieguez, C. M. Mizuno, R. Ghai, F. Rodriguez-Valera, and C. Amaro, "Replicating phages in the epidermal mucosa of the eel (Anguilla anguilla)," *Frontiers Microbiol.*, vol. 6, p. 3, Jan. 2015.

[8] J. A. Dill, A. C. Camus, J. H. Leary, F. Di Giallonardo, E. C. Holmes, and T. F. Ng, "Distinct viral lineages from fish and amphibians reveal the complex evolutionary history of hepadnaviruses," *J. Virol.*, vol. 90, no. 17, pp. 7920–7933, Sep. 2016.

[9] S. Temmam, S. Monteilbouchard, M. Sambou, M. Aubadieladrix, S. Azza, P. Decloquement, J. Y. B. Khalil, J. Baudoin, P. Jardot, C. Robert, B. La Scola, O. Mediannikov, D. Raoult, and C. Desnues, "Faustovirus-like asfarvirus in hematophagous biting midges and their vertebrate hosts," *Frontiers Microbiol.*, vol. 6, p. 1406, Dec. 2015.

[10] M. J. Roossinck, "Deep sequencing for discovery and evolutionary analysis of plant viruses," *Virus Res.*, vol. 239, pp. 82–86, Jul. 2017.

[11] K. E. Wommack, D. J. Nasko, J. Chopyk, and E. G. Sakowski, "Counts and sequences, observations that continue to change our understanding of viruses in nature," *J. Microbiol.*, vol. 53, no. 3, pp. 181–192, Mar. 2015.

[12] J. R. Brum and M. B. Sullivan, "Rising to the challenge: Accelerated pace of discovery transforms marine virology," *Nature Rev. Microbiol.*, vol. 13, no. 3, pp. 147–159, Mar. 2015.

[13] F. H. Coutinho, G. B. Gregoracci, J. M. Walter, C. C. Thompson, and F. L. Thompson, "Metagenomics sheds light on the ecology of marine microbes and their viruses," *Trends Microbiol.*, vol. 26, no. 11, pp. 955–965, Nov. 2018.

[14] E. Luo, F. O. Aylward, D. R. Mende, and E. F. DeLong, "Bacteriophage distributions and temporal variability in the ocean's interior," *MBio*, vol. 8, no. 6, pp. e01903-1–e01903-17, Nov. 2017.

[15] M. Mohiuddin and H. E. Schellhorn, "Spatial and temporal dynamics of virus occurrence in two freshwater lakes captured through metagenomic analysis," *Frontiers Microbiol.*, vol. 6, p. 960, Sep. 2015.

[16] B. Bolduc, J. F. Wirth, A. Mazurie, and M. J. Young, "Viral assemblage composition in Yellowstone acidic hot springs assessed by network analysis," *ISME J.*, vol. 9, no. 10, pp. 2162–2177, Oct. 2015.

[17] A. A. Pratama and J. D. van Elsas, "The 'neglected' soil virome–potential role and impact," *Trends Microbiol.*, vol. 26, no. 8, pp. 649–662, Aug. 2018.

[18] G. Trubl, S. Roux, N. Solonenko, Y. Li, B. Bolduc, J. Rodriguezramos, E. A. Eloefadrosh, V. I. Rich, and M. B. Sullivan, "Towards optimized viral metagenomes for double-stranded and single-stranded DNA viruses from challenging soils," *PeerJ*, vol. 7, p. e7265, Jul. 2019.

[19] L. L. Han, D. T. Yu, L. M. Zhang, J. P. Shen, and J. Z. He, "Genetic and functional diversity of ubiquitous DNA viruses in selected Chinese agricultural soils," *Sci. Rep.*, vol. 7, Mar. 2017, Art. no. 45142.

[20] Y. Wang, X. Jiang, L. Liu, B. Li, and T. Zhang, "High-resolution temporal and spatial patterns of virome in wastewater treatment systems," *Environ. Sci. Technol.*, vol. 52, no. 18, pp. 10337–10346, Sep. 2018.

[21] D. E. Holmes, L. Giloteaux, A. K. Chaurasia, K. H. Williams, B. Luef, M. J. Wilkins, K. C. Wrighton, C. A. Thompson, L. R. Comolli, and D. R. Lovley, "Evidence of geobacter-associated phage in a uranium-contaminated aquifer," *ISME J.*, vol. 9, no. 2, pp. 333–346, Feb. 2015.

[22] A. G. J. Wooley and I. Friedberg, "A primer on metagenomics," *PLoS Comput. Biol.*, vol. 6, no. 2, 2010, Art. no. e1000667.

[23] J. M. Norman, S. A. Handley, M. T. Baldridge, L. Droit, C. Y. Liu, B. C. Keller, A. Kambal, C. L. Monaco, G. Zhao, P. Fleshner, T. S. Stappenbeck, D. P. B. McGovern, A. Keshavarzian, E. A. Mutlu, J. Sauk, D. Gevers, R. J. Xavier, D. Wang, M. Parkes, and H. W. Virgin, "Disease-specific alterations in the enteric virome in inflammatory bowel disease," *Cell*, vol. 160, no. 3, pp. 447–460, Jan. 2015.

[24] F. Rohwer and R. V. Thurber, "Viruses manipulate the marine environment," *Nature*, vol. 459, no. 7244, pp. 207–212, May 2009.

[25] S. Mills, F. Shanahan, C. Stanton, C. Hill, A. Coffey, and R. P. Ross, "Movers and shakers: Influence of bacteriophages in shaping the mammalian gut microbiota," *Gut Microbes*, vol. 4, no. 1, pp. 4–16, Jan./Feb. 2013.

[26] S. R. Modi, H. H. Lee, C. S. Spina, and J. J. Collins, "Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome," *Nature*, vol. 499, no. 7457, pp. 219–222, Jul. 2013.

[27] S. R. Abeles and D. T. Pride, "Molecular bases and role of viruses in the human microbiome," *J. Mol. Biol.*, vol. 426, no. 23, pp. 3892–3906, Nov. 2014.

[28] C. Rinke *et al.*, "Insights into the phylogeny and coding potential of microbial dark matter," *Nature*, vol. 499, no. 7459, pp. 431–437, Jul. 2013.

[29] Y. Li, H. Wang, K. Nie, C. Zhang, Y. Zhang, J. Wang, P. Niu, and X. Ma, "VIP: An integrated pipeline for metagenomics of virus identification and discovery," *Sci. Rep.*, vol. 6, Mar. 2016, Art. no. 23774.

[30] G. Zhao, G. Wu, E. S. Lim, L. Droit, S. R. Krishnamurthy, D. H. Barouch, H. W. Virgin, and D. Wang, "VirusSeeker, a computational pipeline for virus discovery and virome composition analysis," *Virology*, vol. 503, pp. 21–30, Mar. 2017.

[31] S. Rampelli, M. Soverini, S. Turroni, S. Quercia, E. Biagi, P. Brigidi, and M. Candela, "ViromeScan: A new tool for metagenomic viral community profiling," *BMC Genomics*, vol. 17, p. 165, Mar. 2016.

[32] S. S. Tithi, F. O. Aylward, R. V. Jensen, and L. Zhang, "FastVirome-Explorer: A pipeline for virus and phage identification and abundance profiling in metagenomics data," *PeerJ*, vol. 6, p. e4227, Jan. 2018.
[33] S. Roux, F. Enault, B. L. Hurwitz, and M. B. Sullivan, "VirSorter: Mining viral signal from microbial genomic data," *PeerJ*, vol. 3, p. e985, May 2015.
[34] A. Garretto, T. Hatzopoulos, and C. Putonti, "virMine: Automated detection of viral sequences from complex metagenomic samples," *PeerJ*, vol. 7, p. e6695, Apr. 2019.
[35] J. Ren, N. A. Ahlgren, Y. Y. Lu, J. A. Fuhrman, and F. Sun, "VirFinder: A novel k-mer based tool for identifying viral sequences from assembled metagenomic data," *Microbiome*, vol. 5, no. 1, p. 69, Jul. 2017.
[36] T. Zheng, J. Li, Y. Ni, K. Kang, M. Misiakou, L. Imamovic, B. K. C. Chow, A. A. Rode, P. Bytzer, M. O. A. Sommer, and G. Panagiotou, "Mining, analyzing, and integrating viral signals from metagenomic data," *Microbiome*, vol. 7, no. 1, p. 42, Mar. 2019.
[37] I. Borozan, S. Watt, and V. Ferretti, "Integrating alignment-based and alignment-free sequence similarity measures for biological sequence classification," *Bioinformatics*, vol. 31, no. 9, pp. 1396–1404, May 2015.
[38] D. Amgarten, L. P. P. Braga, A. M. da Silva, and J. C. Setubal, "MARVEL, a tool for prediction of bacteriophage sequences in metagenomic bins," *Frontiers Genet.*, vol. 9, p. 304, Aug. 2018.
[39] J. Droge and A. C. McHardy, "Taxonomic binning of metagenome samples generated by next-generation sequencing technologies," *Brief Bioinf.*, vol. 13, no. 6, pp. 646–655, Nov. 2012.
[40] D. D. Kang, J. Froula, R. Egan, and Z. Wang, "MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities," *PeerJ*, vol. 3, p. e1165, Aug. 2015.
[41] S. Akhter, R. K. Aziz, and R. A. Edwards, "PhiSpy: A novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies," *Nucleic Acids Res.*, vol. 40, no. 16, p. e126, Sep. 2012.
[42] G. Yu, Y. Jiang, J. Wang, H. Zhang, and H. Luo, "BMC3C: Binning metagenomic contigs using codon usage, sequence composition and read coverage," *Bioinformatics*, vol. 34, no. 24, pp. 4172–4179, Dec. 2018.
[43] A. L. Grazziotin, E. V. Koonin, and D. M. Kristensen, "Prokaryotic virus orthologous groups (pVOGs): A resource for comparative genomics and protein family annotation," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D491–D498, Jan. 2017.
[44] M. Y. Galperin, K. S. Makarova, Y. I. Wolf, and E. V. Koonin, "Expanded microbial genome coverage and improved protein family annotation in the COG database," *Nucleic Acids Res.*, vol. 43, no. D1, pp. D261–D269, Jan. 2015.
[45] L. A. Ogilvie, L. D. Bowler, J. Caplin, C. Dedi, D. Diston, E. Cheek, H. Taylor, J. Ebdon, and B. V. Jones, "Genome signature-based dissection of human gut metagenomes to extract subliminal viral sequences," *Nature Commun.*, vol. 4, Sep. 2013, Art. no. 2420.
[46] B. M. Satinsky, B. L. Zielinski, M. Doherty, C. B. Smith, S. Sharma, J. H. Paul, B. C. Crump, and M. A. Moran, "The Amazon continuum dataset: Quantitative metagenomic and metatranscriptomic inventories of the Amazon River plume, June 2010," *Mbio*, vol. 2, no. 1, p. 17, 2014.
[47] S. Boisvert, F. Raymond, E. Godzaridis, F. Laviolette, and J. Corbeil, "Ray meta: Scalable de novo metagenome assembly and profiling," *Genome Biol.*, vol. 13, no. 12, pp. 1–13, 2012.
[48] Y. Y. Lu, T. Chen, J. A. Fuhrman, and F. Sun, "COCACOLA: Binning metagenomic contigs using sequence COmposition, read CoverAge, CO-alignment and paired-end read LinkAge," *Bioinformatics*, vol. 33, no. 6, pp. 791–798, 2016.
[49] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *J. Mol. Biol.*, vol. 215, no. 3, pp. 403–410, Oct. 1990.
[50] B. L. Hurwitz, A. Ponsero, J. Thornton, and J. M. U'Ren, "Phage hunters: Computational strategies for finding phages in large-scale 'omics datasets," *Virus Res.*, vol. 244, pp. 110–115, Jan. 2018.

QIAOLIANG LIU received the B.S. and M.S. degrees from the College of Computer Science and Technology, Jilin University, where he is currently pursuing the Ph.D. degree with the Laboratory of Pattern Recognition and Artificial Intelligence, College of Communication and Engineering. His research areas include machine learning and bioinformatics.

FU LIU received the B.S. and M.S. degrees from the Jilin University of Technology, in 1991 and 1994, respectively, and the Ph.D. degree from the Department of Control Science and Engineering, Jilin University, in 2002. He is currently a Professor with Jilin University. His research interests include machine vision, pattern recognition, bioinformatics, and biometrics.

JIAXUE HE received the joint Ph.D. degree from the Division of Orthopedics, Department for Clinical Science, Intervention and Technology (CLINTEC), Karolinska Institute, Sweden, in 2012, and the Ph.D. degree from the Norman Bethune College of Medicine, Jilin University, in 2013. She was a Lab Assistant with the Division of Orthopedics, Department of CLINTEC, Karolinska Institute. She is currently a Technician with the Genetic Diagnosis Center, The First Hospital of Jilin University. Her area of research includes the sequencing and analysis of tumor and microbiology genes.

MIAOLEI ZHOU (M'12) was born in 1976. He received the B.S. and M.S. degrees in industrial electric automation from the Jilin Institute of Technology, China, in 1997 and 2000, respectively, and the Ph.D. degree in control theory and control engineering from Jilin University, China, in 2004.

From 2006 to 2008, he was a Postdoctoral Researcher with Tokyo University, Japan. In 2000, he joined the Department of Control Science and Engineering, Jilin University. He became an Associate Professor, in 2009, and a Professor, in 2014. He has supervised over 20 research projects, including The National Natural Science Funds of China and National High Technology Research and Development Program. He is the author of more than 80 articles. His research interests include micro/nano drive and control technology, nonlinear control theory, and navigation and control of robot.

TAO HOU received the B.S. degree from the College of Mathematics, and the M.S. and Ph.D. degrees from the College of Communication and Engineering, Jilin University. She is currently a Lecturer with Jilin University. Her areas of research include machine learning and bioinformatics.

YUN LIU (M'18) received the B.S. and Ph.D. degrees from the College of Communication and Engineering, Jilin University, in 2011 and 2016, respectively. He is currently a Lecturer with Jilin University. His areas of research include machine learning and bioinformatics.