MLR-OOD: a Markov chain based Likelihood Ratio method for Out-Of-Distribution detection of genomic sequences

Xin Bai[1], Jie Ren[2], and Fengzhu Sun[1,*]

[1]Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA, 90089, USA
[2]Google Research, Brain Team, USA
[*]Corresponding author: Fengzhu Sun, fsun@usc.edu

Abstract

Machine learning or deep learning models have been widely used for taxonomic classification of metagenomic sequences and many studies reported high classification accuracy. Such models are usually trained based on sequences in several training classes in hope of accurately classifying unknown sequences into these classes. However, when deploying the classification models on real testing data sets, sequences that do not belong to any of the training classes may be present and are falsely assigned to one of the training classes with high confidence. Such sequences are referred to as out-of-distribution (OOD) sequences and are ubiquitous in metagenomic studies. To address this problem, we develop a deep generative model-based method, MLR-OOD, that measures the probability of a testing sequencing belonging to OOD by the likelihood ratio of the maximum of the in-distribution (ID) class conditional likelihoods and the Markov chain likelihood of the testing sequence measuring the sequence complexity. We compose three different microbial data sets consisting of bacterial, viral, and plasmid sequences for comprehensively benchmarking OOD detection methods. We show that MLR-OOD achieves the state-of-the-art performance demonstrating the generality of MLR-OOD to various types of microbial data sets. It is also shown that MLR-OOD is robust to the GC content, which is a major confounding effect for OOD detection of genomic sequences. In conclusion, MLR-OOD will greatly reduce false positives caused by OOD sequences in metagenomic sequence classification.

1 Introduction

Classification of metagenomic sequences into different taxons is an essential problem for understanding the compositions of microbial communities. Some mapping based computational software tools [1, 2, 3, 4, 5, 6] can rapidly and accurately classify mappable microbial sequences based on reference databases. However, due to the lack of a complete set of reference genomes, such mapping based approaches are not able to discover novel sequences in specific classes, which is becoming increasingly important since a large portion of the metagenomic sequences, the so-called "microbial dark matter", remain poorly understood [7, 8, 9]. For example, several studies have estimated that around 85%-99% of bacteria and archaea cannot be cultured [8] and up to 60%-80% bacterial sequences in some environments belong to unknown taxons [10, 11, 12]. Meanwhile, only several thousand of viral species have been recognized by 2017 [13]. Recently, many machine learning or deep learning based approaches have been developed to classify metagenomic sequences into several classes without depending on reference databases [14, 15, 16, 17, 18]. These approaches are usually based on classification models trained on sequences in several training classes. High classification accuracies have been reported in these studies and it is believed that these approaches can generalize well to unknown sequences, meaning that novel sequences belonging to these classes can also be discovered. However, due to the complexity of compositions of metagenomic sequences, it is inevitable for these approaches to deal with sequences that do not belong to any of the training classes, the so-called out-of-distribution (OOD) genomic sequences. Although microbial researchers usually tend to believe that these OOD sequences will receive low classification scores in real applications, this assumption is unlikely to hold true as some studies have reported

that modern neural network classifiers may assign higher classification scores for OOD inputs [19, 20, 21]. Thus, the detection of OOD genomic sequences is urgently needed to ensure that only in-distribution (ID) sequences, i.e. those belonging to one of the training classes, will be classified so that the credibility of taxonomic classification approaches can be improved.

The detection of OOD inputs is an active research topic in machine learning and many approaches have been proposed to address this problem [22, 23, 24, 25, 26, 27, 28, 29]. However, most of the current methods, either generative model based or discriminative model based, are designed for detecting OOD images on which very high prediction accuracy can be easily achieved. However, image data and genomic sequence data are distinct in nature. For example, genomic sequences are one-dimensional data with only four possible nucleotides at each position, while images are multi-dimensional data with much more complex pixel values for different dimensions of colors. Therefore, most of these methods are either not directly applicable or have much lower prediction accuracy on the genomic data [28]. Among all these attempts, the likelihood ratio (LLR) method proposed by Ren et al. particularly addressed the problem of detecting OOD genomic sequences [28]. The LLR method uses likelihood ratios of two generative models for the ID data and randomly perturbed ID data, based on the hypothesis that only the semantic part of a sequence is associated with its taxon while the background part, in contrast, bias the prediction of OOD sequences. For example, the semantic part can be understood as the coding genes or motifs specific to a taxon and the background part can be understood as the repeated regions or transferred genes shared by sequences in different taxons. Ren et al. [28] compared LLR with nine different approaches and showed that their approach achieves the highest prediction accuracy on detecting OOD bacterial sequences. It is also shown in [28] that LLR is robust to the GC content of genomic sequences which is a major confounding effect on detecting OOD genomic sequences. Another major contribution of LLR is that it composed the first data set consisting of bacterial sequences from multiple genera for benchmarking future OOD detection methods.

Despite its effectiveness, the prediction accuracy of LLR for detecting OOD genomic sequences is still not very high, leaving room for improvement based on more powerful methods. For example, the reported optimal area under the receiver operating characteristic curve (AUROC) for predicting 250bp bacterial OOD sequences using LLR is 0.755 [28], which is based on tuning two model-parameters. Besides, LLR requires a validation data set containing both ID and OOD sequences for model-parameters tuning, leading to the concern that the model trained on one data set may not be able to easily generalize to another data set. For example, Ren et al. [28] used sequences in 10 bacterial genera as the ID sequences and constructed the OOD validation data set using OOD sequences from 60 other bacterial genera. However, in practice, we usually have knowledge on the ID sequences only and OOD sequences can belong to any other classes such as sequences in other bacterial genera, viral sequences, contaminations from the human genome, etc. Therefore, it is unrealistic to expect that the model-parameters of LLR tuned on a specific validation data set can perform well on any other testing data set.

In this paper, we present MLR-OOD, a Markov chain based likelihood ratio method, for detecting OOD genomic sequences. MLR-OOD detects OOD genomic sequences based on the likelihood ratio of the maximum of the ID class conditional likelihood and the likelihood of the sequence under a Markov chain model mimicking the sequence complexity of testing sequences. The rationale of MLR-OOD is based on two aspects. First, compared to LLR that uses one general model for all ID sequences, MLR-OOD uses the maximum of the ID class conditional likelihood taking into account different models for sequences in various ID classes, promoting a more precise modeling of the ID sequences. Second, based on the assumption that input sequence complexity, which can be modeled by Markov chain likelihood, is a factor that biases OOD detection, we propose to use Markov chain likelihoods to adjust the maximum of the ID class conditional likelihood, bypassing the generation of background null sequences used by LLR. Thus, MLR-OOD is completely free of tuning model-parameters. On the other hand, LLR depends on the optimal perturbation rate that needs to be determined by a validation set consisting of both ID and OOD sequences. In addition to the bacterial data sets composed by Ren et al. [28], we composed two more microbial data sets for viruses and plasmids, to more comprehensively benchmark the performance of MLR-OOD. We show that MLR-OOD yields notably higher prediction accuracy than LLR. We also conclude that MLR-OOD is robust to the GC content on almost all data sets while LLR can be somewhat biased on the viral and plasmid data sets.

In summary, we have two major contributions in this paper. First, we construct viral and plasmid data sets that can be jointly used with the bacterial data set composed by Ren et al. [28] for comprehensively benchmarking the performances of different OOD detection methods for genomic sequences. Second, we

Table 1: The data sets we use for benchmarking the detction of OOD genomic sequences. The details about the specific ID and OOD genera names and the construction of the training and testing data sets are discussed in Supplementary Materials.

Type	Bacteria		Virus	Plasmid
Name	**Test2016**	**Test2018**	N/A	N/A
Phytogenetic level for classes	Genus		Family	Genus
Number of ID classes	10		6	6
Number of OOD classes	60		20	20
Description	The same data set as the one used in [28]	The same genera as the **Test2016** data set, but more novel	Viruses whose hosts are bacteria and archaea	Plasmids whose hosts are bacteria and archaea

develop MLR-OOD, a powerful method achieving the current state-of-the-art prediction accuracy for OOD detection of genomic sequences without tuning model-parameters on validation data sets. We believe that MLR-OOD will improve the credibility of machine learning based metagenomic taxonomic classification.

2 Materials and Methods

2.1 Data sets for detecting OOD genomic sequences

Ren et al. [28] composed a genomic sequence data set consisting of bacterial sequences in different genera that can serve as a benchmark for detecting OOD genomic sequences. Although bacteria are abundantly distributed in microbial communities, there are many other types of molecules including viruses, plasmids, fungi, archaea, etc [30, 31]. To benchmark the prediction accuracy of OOD detection on different types of molecules, we construct a new testing data set for bacteria and two more data sets for viruses and plasmids, respectively. A brief summary of the data sets we use is shown in Table 1.

2.1.1 The bacterial data set

Ren et al. have already constructed a comprehensive bacterial data set that is publicly available in [28] for OOD detection. We use the same data set and add another new testing data set to demonstrate the prediction accuracy of MLR-OOD.

Specifically, Ren et al. downloaded 11,672 bacteria genomes from National Center for Biotechnology Information (NCBI) and chopped the genomes to short 250bp sequences [28]. Different genera of these genomes were used to define the class labels (both ID and OOD classes). For example, genomes in 10 particular bacterial genera and that were discovered before 01/01/2011 are used to construct the ID training data set. The validation data set contained ID sequences in these 10 genera but were discovered between 01/01/2011 and 01/01/2016, and OOD sequences in other 60 genera discovered in the same time period. The testing data set similarly contained ID sequences in the 10 genera and that were discovered between 01/01/2016 and September 2018, and OOD sequences in the 60 genera that did not overlap with the validation OOD genera discovered in the same time period. We collected the accession numbers of the original bacterial genomes and downloaded them from NCBI in order to chop the bacterial genomes into non-overlapping sequence fragments of different training and testing lengths as discussed in Sections 2.5 and 2.6. We choose non-overlapping training and testing sequences to avoid potentially redundancy information from overlapping sequences. To distinguish from our newly constructed testing data set, we refer to the testing data set consisting of bacterial genomes used by Ren et al. as the **Test2016** data set. For the details, please refer to [28].

We also constructed another testing data set consisting of ID and OOD sequences discovered between 10/01/2018 and 10/1/2021 to benchmark OOD detection methods on relatively more novel sequences. The ID and OOD sequences were from the same ID and OOD genera used by Ren et al. [28] for constructing the **Test2016** data set. We refer to the new testing data set as the **Test2018** data set.

2.1.2 The viral data set

We downloaded 1295 viral genomes whose hosts are bacteria and archaea from NCBI and then constructed an ID training data set and a testing data set containing both ID and OOD sequences. Since viral sequences from NCBI are much fewer and shorter compared to bacterial sequences, we used different viral families to define ID and OOD classes. Viral genomes that were in 6 particular families and were discovered before 01/01/2016 were used to construct 6 ID viral classes. Viral genomes that were in these families but were discovered between 01/01/2016 and 10/01/2021 were treated as ID testing sequences. The OOD testing data set contained viral genomes in 20 other randomly chosen families and were also discovered between 01/01/2016 and 10/01/2021. Since potentially highly similar sequences will bias the training and testing process, we use CD-HIT [32] with parameters "−c 0.95 −n 10 −M 0 −T 16" to cluster all the sequences and then remove the duplicate ones. The viral genomes were then chopped into non-overlapping sequence fragments of length 250bp for training and various lengths for testing. The details about the specific ID and OOD family names and the construction of the training and testing data sets are discussed in Supplementary Materials.

2.1.3 The plasmid data set

We also downloaded 818 plasmid genomes whose hosts are bacteria and archaea from NCBI to build up a data set for detecting OOD plasmid sequences. We used plasmid genera to define ID and OOD classes, the same as the bacterial data set. Similar to Section 2.1.2, we built up an ID training data set containing plasmid genomes discovered before 01/01/2016 in 6 different classes. The testing data set contains plasmid genomes discovered between 01/01/2016 and 10/01/2021 in 6 ID and 20 randomly chosen OOD classes. The plasmid genomes were similarly clustered by CD-HIT and then chopped into non-overlapping sequence fragments of length 250bp for training and various lengths for testing. The details about the specific ID and OOD genera names and the construction of the training and testing data sets are discussed in Supplementary Materials.

2.2 Outline of the Methods

Suppose we have an in-distribution (ID) data set of genomic sequences $D(\mathcal{X}, \mathcal{Y})$. Let the pair $(\mathbf{x}, y)$ denote an individual nucleotide sequence $\mathbf{x} = x_1 \dots x_d \dots x_D, \mathbf{x} \in \mathcal{X}, x_d \in \{A, C, G, T\}$ and its in-distribution class label $y \in \mathcal{Y} := \{1, \dots, k, \dots, K\}$), i.e., the specific taxon that sequence belongs to, where K denotes the number of in-distribution classes and D denotes the sequence length. We are interested in these ID sequences belonging to particular taxons and aim to detect other sequences that do not belong to any of the K ID classes, i.e. $y \notin \mathcal{Y}$, that are referred to as the OOD sequences, for accurate downstream analyses of both ID and OOD sequences.

We develop MLR-OOD, a Markov chain based likelihood ratio method combining both the ID generative models and the Markov models for the testing sequences, for the detection of OOD genomic sequences. Before discussing the details of our method, we first give a brief review on the LLR method proposed by Ren et al. [28] which was the first study paying particular attention to detecting OOD genomic sequences.

2.3 The framework of the likelihood ratio method

The LLR method proposed by Ren et al. uses the likelihood ratio of an original model and a background model to measure the chance of being OOD for the testing sequences [28]. The original model, denoted as $p_\theta(\cdot)$, is a generative model trained on all the ID sequences that can be naturally used to measure the likelihoods of being OOD for testing sequences. However, Ren et al. observed that the likelihood of the original model fails to separate ID and OOD testing sequences and is biased by the GC content of each testing sequence [28]. This observation can be explained by the fact that the original model captures both the semantic and the background parts of the ID sequences but only the semantic part is helpful for OOD detection. The GC content is considered a confounding effect arising from the background part of the sequences, thus motivating them to generate a background model that can adjust the confounding effects of the original model. Based on the assumption that random perturbations can corrupt the semantic part in the data, Ren et al. [28] randomly perturbed a certain fraction μ of the nucleotides in the original sequences by

changing the specific nucleotides to the other three with equal probability to build a null set of background sequences. The background model $p_{\theta_0}(\cdot)$ is then trained on the background sequences. For an incoming testing sequence $\mathbf{x}$, the predicting score $S_{\text{LLR}}(\mathbf{x})$ is calculated as the log-likelihood ratio of the original and background models

$$S_{\text{LLR}}(\mathbf{x}) = \log \frac{p_\theta(\mathbf{x})}{p_{\theta_0}(\mathbf{x})} = \log p_\theta(\mathbf{x}) - \log p_{\theta_0}(\mathbf{x}). \tag{1}$$

As shown in equation (1), the information of the background part, which is contained in both the original and background likelihoods, is canceled out by taking the ratio, so that the semantic part in the original model stands out for OOD detection. A larger value of $S_{\text{LLR}}(\mathbf{x})$ indicates a higher chance of being ID (equivalently, a lower chance of being OOD) for sequence $\mathbf{x}$.

The original and background models were trained using long short-term memory (LSTM) [33], a deep generative model that has been widely used to model genomic sequences [34, 35, 36], on entire original or background sequences. Specifically, Ren et al. [28] used one-hot encoder to transform the genomic sequences from strings comprised of $\{A, C, G, T\}$ nucleotides to binary numeric vectors and then feed them to a LSTM layer, followed by a dense layer and a softmax function to predict the probability distribution over $\{A, C, G, T\}$ at each position. The LSTM model was jointly trained at all positions from 1 to D. Then, if we denote the LSTM model trained for the original model by $p_\theta(\mathbf{x})$, the log-likelihood $\log p_\theta(\mathbf{x})$ for any incoming testing sequence $\mathbf{x}$ can be calculated as follows

$$\log p_\theta(\mathbf{x}) = \sum_{d=1}^{D} \log p_{k,\theta}(x_d | x_{<d}),$$

where $x_{<d}$ means all the nucleotides of sequence $\mathbf{x}$ before position d. The calculation for the log-likelihood of the background model $\log p_{\theta_0}(\mathbf{x})$ is similar as above.

Two model-parameters need to be carefully tuned during the training process of LLR: the perturbation rate μ and the coefficient of L_2 regularization λ added to the model weights when training the background model (optional). Ren et al. used a validation data set consisting of additional ID and OOD data to determine these two model-parameters [28].

2.4 The framework of the new MLR-OOD method

Although the LLR method achieves a higher prediction accuracy and is more robust to the GC content compared to the method only using the original likelihood, there is still room for improvement in two aspects. First, the LLR method only considers a general model $p_\theta(\cdot)$ for all ID sequences and does not consider the K ID classes when modeling ID training data. This may not be optimal since sequences in different ID classes may follow different models. Therefore, we borrow the ideas of using likelihood ratios from [28] but utilize the information of the ID classes to further increase the prediction accuracy. Second, the LLR method relies on the validation data set for tuning model-parameters, that is, to have access to some of the OOD data, which is unlikely to remain true in real practice.

We first propose using the the likelihood that is maximum across LSTM models across all ID classes $p_{\max,\theta}(\cdot)$ instead of $p_\theta(\cdot)$ for modeling the ID data. The high level idea is to train the generative models for the data in each ID class separately and choose the most appropriate model for new testing sequences by taking the maximum of the class conditional likelihoods across all ID classes. In addition to using LSTM which is also used by LLR to train the generative models, we also tried Markov chains which are generally more interpretable than LSTM on the bacterial data set but the prediction accuracy was much lower, at AUROC around 0.6 for 250bp bacterial sequences. We present the details and the results in Supplementary Materials. Therefore, we only present our LSTM model and the corresponding results in the rest of the main text. Let $p_{k,\theta}(\cdot)$ be the model we trained on the sequences in the k-th ID class. For each incoming sequence $\mathbf{x}$, we calculate the maximum of the class conditional likelihoods as follows

$$p_{\max,\theta}(\mathbf{x}) \quad = \quad \max_{k \in \{1,\cdots,K\}} \; p_{\theta_k}(\mathbf{x}), \tag{2}$$

where $p_{\theta_k}(\mathbf{x})$ denotes the LSTM likelihood of sequence $\mathbf{x}$ belonging to class k. We use the same model parameters as in [28] for training the LSTM models on the ID data for the methods listed above. In detail,

the size of the hidden units in the LSTM model is 2,000. The number of epoches is 900,000 and the learning rate is 0.0005. The batch size is 100 and Adam optimizer is used to minimize the training loss. We denote the method using $p_{\max,\theta}(\mathbf{x})$ as the prediction score by the **max-LL** method. As a preliminary method for predicting OOD sequences proposed in this paper, we will show in Section 3 that the max-LL method itself is a very powerful predicting statistic compared to S_{LLR}, even without background adjustment. Note that the mixture model likelihood $p_{\text{mix},\theta}(\cdot) = \sum_{k=1}^{K} p_{\theta_k}(\mathbf{x})p(k)$ is another natural way to more precisely model the ID data, where $p(k)$ denotes the prior probability for each ID class. However, it is challenging to estimate $p(k)$ for real metagenomic data since a large portion of the metagenomic sequences still cannot be classified. Therefore, we adopt $p_{\max,\theta}(\mathbf{x})$ for our method.

However, since the prediction score of the max-LL method is still likely to be biased by confounding effects such as the GC content, another key question is yet remaining: how to adjust $p_{\max,\theta}(\mathbf{x})$ to bypass the effect of confounding effects such that the prediction accuracy can be further increased without tuning model-parameters? We propose to adjust the maximum of the class conditional likelihoods $p_{\max,\theta}(\mathbf{x})$ by the Markov chain likelihood of the testing sequence which can be regarded as a special case of sequence complexity. Serra et al. observed that input complexity can bias the likelihood as relatively "simple" patterns generally have higher likelihoods compared to complex patterns [37]. Several measures have been proposed [38, 39] to model the complexity of genomic sequences. Among all these sequence complexity measures, the CE complexity [39] based on entropy of sequences is essentially a constant times the log-likelihood of the testing sequence under the independent and identically distributed (i.i.d.) model, which is equivalently, a zero-th order Markov chain. The GC content is essentially related to the i.i.d. likelihood if we do not distinguish between G and C (also A and T) nucleotides. Inspired by the idea that the likelihood under the i.i.d. model may confound OOD detection, we generalize the idea and use the log-likelihood of each testing sequence modeled by a Markov chain to adjust $p_{\max,\theta}(\mathbf{x})$ given that Markov chains have been widely used in genomic sequence analyses [40, 41, 42, 43, 44, 45, 46, 47]. Specifically, assume that the testing sequence $\mathbf{x} = x_1 \dots x_d \dots x_D$ is modeled by a r-th order Markov chain, then the log-likelihood of $\mathbf{x}$ is estimated as follows if we safely disregard the initial distribution $\pi(x_1 \cdots x_r)$

$$\begin{aligned} L^r_{\text{MC}}(\mathbf{x}) &= \log \pi(x_1 \cdots x_r) + \sum_{\mathbf{w}} N_{\mathbf{w}} \log P(w_{r+1}|\mathbf{w}-), \\ &\approx \sum_{\mathbf{w}} N_{\mathbf{w}} \log \frac{N_{\mathbf{w}}}{N_{\mathbf{w}-}}, \end{aligned}$$

where $\mathbf{w} = w_1 w_2 \cdots w_{r+1}$ denotes a word of nucleotides of length $r+1$, $N_{\mathbf{w}}$ denotes the count of occurrences of $\mathbf{w}$ in sequence $\mathbf{x}$, $\mathbf{w}- = w_1 w_2 \cdots w_r$ denotes word $\mathbf{w}$ with the last letter removed, and $N_{\mathbf{w}-}$ is the count of occurrences of $\mathbf{w}-$. When $r = 0$, $N_{\mathbf{w}-}$ degenerates to the sequence length D. The transition probability is denoted by $P(w_{r+1}|\mathbf{w}-)$ and is estimated by $\frac{N_{\mathbf{w}}}{N_{\mathbf{w}-}}$ using the maximum likelihood estimation.

Therefore, we propose to use $S^r_{\text{MLR-OOD}}(\mathbf{x})$ defined in equation (3) for OOD genomic sequences detection.

$$S^r_{\text{MLR-OOD}}(\mathbf{x}) = \log p_{\max,\theta}(\mathbf{x}) - L^r_{\text{MC}}(\mathbf{x}). \tag{3}$$

A larger value of $S^r_{\text{MLR-OOD}}(\mathbf{x})$ indicates a higher probability of belonging to ID. Note that the calculation of $S^r_{\text{MLR-OOD}}(\mathbf{x})$ is completely free of tuning model-parameters since r is data-driven, determined by the most commonly estimated Markov order of the testing sequences, and has nothing to do with the training process. We use Bayesian information criterion (BIC) [48] to estimate the Markov orders of the testing sequences.

To facilitate the understanding of MLR-OOD, we present its complete workflow in Figure 1 summarizing the above procedures.

2.5 Investigating the effect of training sequence length on prediction accuracy and computational time

Ren et al. [28] chopped the bacterial genomes into short sequences of 250bp for training the LSTM model. However, the effect of the training sequence length on prediction accuracy is yet unknown since different lengths of training sequences may lead to different model performances and computational time. We study the effect of training sequence length on our MLR-OOD method by chopping the original bacterial genomes used by [28] into short sequences of 100bp, 250bp, and 500bp, respectively. Then we train the LSTM models

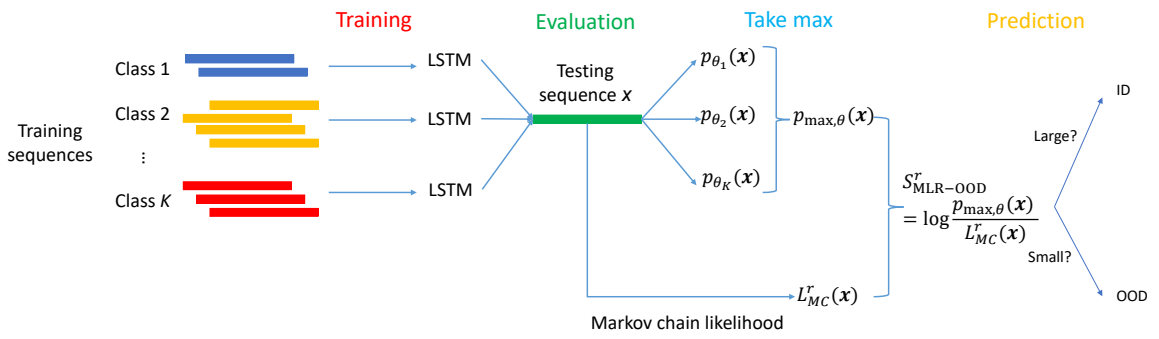

Figure 1: The complete workflow of MLR-OOD for predicting OOD sequences.

for ID classes based on different training lengths and test the corresponding model performances on the **Test2016** bacterial testing data set with testing contig lengths being 1000bp, 2500bp, and 5000bp. We follow the protocol in Section 2.6 to deal with different training and testing lengths. Our target is to select the best training sequence length for MLR-OOD balancing both prediction accuracy and computational time. We will use the selected training sequence length on the other data sets and compare with the LLR method.

2.6 Investigating the effect of testing contig length on prediction accuracy

Ren et al. present the usefulness of the LLR method by showing the prediction accuracy on 250bp testing sequences [28]. However, in reality, metagenomic reads are usually assembled from short reads of several hundred of basepairs into contigs that are consecutive regions of genomes with overlapping reads to facilitate downstream analyses. These contigs usually have various lengths greater than 250bp. To show the effect of testing contig length on the prediction accuracy of MLR-OOD, we fix the training sequence length as 250bp as it performs the best compared to 100bp and 500bp as shown in Figure 3 and chop the testing genomes into contigs of lengths 250bp, 500bp, 1,000bp, 2,500bp, and 5,000bp. For each contig length except 2,500bp and 5,000bp for the bacterial **Test2018** and the viral data set, 10k ID and 10k OOD contigs are randomly selected to build up the corresponding testing data sets. The bacterial **Test2018** data set contains 4k OOD 2,500bp and 2k OOD 5,000bp testing contigs. The viral data set contains 5k ID and 5k OOD 2,500bp tesing contigs and 3k ID and 3k OOD 5,000bp testing contigs. It is expected that longer contigs generally contain more information about the taxons and thus should have higher prediction accuracy for OOD detection. To overcome the discrepancy of the training and testing sequence lengths, we follow the ideas in [49] by splitting the testing contigs into non-overlapping fragments of 250bp and calculating the prediction score for each of these fragments. The final predicting score for one contig is calculated by averaging the prediction scores of all fragments in that contig. We follow the same protocol to compare MLR-OOD with other methods.

2.7 Investigating the effect of the genome distance between OOD testing sequences and ID training sequences on prediction accuracy

In real applications, the OOD sequences may come from any genetic materials other than the ID sequences. Therefore, it would be interesting to study how the prediction accuracy of MLR-OOD changes with the overall similarity between OOD testing sequences and our ID training sequences. Ren et al. concluded that the prediction accuracy of LLR becomes generally higher if the minimum d_2^S distance [50, 51] between the OOD bacterial classes and the ID bacterial classes increases. We investigate the same problem on all bacterial, viral, and the plasmid data sets we built for MLR-OOD as a byproduct of our analyses. First, we combine all the genomes in each ID training class into a single fasta file representing that class. Second, for each OOD testing genome, we calculate its pairwise Mash distance [52] to all the ID training classes

and then take the minimum. We specify the number of sketches as "-s 1000000" and use default values for other parameters for decent performance of the Mash distance. The reason for choosing the Mash distance is due to that it achieves much faster computational time than d_2^S. Once we obtain the minimum distance to ID classes for each OOD testing genome, we cut a threshold and then only select those genomes having a minimum distance greater than that threshold to build the OOD testing data set. Since currently most metagenomic contigs have length greater or equal to 1,000bp, we cut the ID and OOD testing genomes to 1,000bp sequences and then calculate the prediction accuracy metrics (see Section 2.10) based on MLR-OOD to study the relationship between the Mash distance threshold and the prediction accuracy. We use the same ID testing data set while varying the OOD testing data set in comparison. The number of available OOD testing sequences after constraining the threshold on the minimum Mash distance is certainly fewer and we report the detailed numbers for each type of sequences in Supplementary Materials.

2.8 Investigating the effect of chimeric contigs on the prediction accuracy of MLR-OOD

Chimeric contigs can happen in the assembly process of metagenomic sequences since multiple closely related species usually exist in metagenomic samples [53, 54]. In practice, OOD detection methods may face chimeric contigs containing sequence fragments from both ID and OOD genomes. We investigate the performance of MLR-OOD when dealing with such chimeric contigs. First, for each given fraction c, we simulate 10k chimeric contigs in which the proportion of nucleotides belonging to ID genomes is c and the proportion of nucleotides belonging to OOD genomes is $1-c$. The contig length is fixed at 1,000bp. For example, if $c = 0.25$, we randomly sample a 250bp sequence fragment from the ID genomes in a particular testing data set, then sample another 750bp sequence fragment from the OOD genomes, and finally insert the 750bp OOD sequence fragment at a random position of the 250bp ID sequence fragment. We test different values of $c =$ 0, 0.25, 0.5, 0.75, and 1 for all bacterial, viral, and plasmid data sets. Second, we calculate the prediction scores of MLR-OOD on those chimeric contigs following the procedures introduced in Section 2.6. Third, we calculate the prediction accuracy metric AUROC (see Section 2.10) for classifying completely OOD contigs ($c = 0$) and contigs containing a certain fraction of ID sequences ($c = 0.25$, 0.5, 0.75, and 1) based on the prediction scores of MLR-OOD. We study how the distribution of the prediction scores and the prediction accuracy change with different values of c.

2.9 Comparison to LLR and other classifier-based methods for OOD detection

Since Ren et al. [28] already compared the LLR method with a large number of other methods [22, 23, 24, 55, 56, 57, 58], most of which are designed for detecting OOD images, and showed that the LLR method achieved the best prediction accuracy, we first focus on comparing our proposed methods with LLR for detecting OOD genomic sequences. In particular, we first comapre LLR with two methods proposed in this paper: 1) the max-LL method directly using the maximum of the in-distribution (ID) class conditional likelihoods, and 2) the refined MLR-OOD method on the bacterial, viral, and plasmid data sets.

Unlike the LLR method, the max-LL and the MLR-OOD methods do not have model-parameters to be tuned. Therefore, we do not need a validation data set. To compare with the LLR method on the bacterial data set, we use the optimal model-parameters $\mu = 0.1$ and $\lambda = 10^{-4}$ that have already been chosen by Ren et al. for the LLR method based on the validation data set [28]. We try different values of the perturbation rate $\mu = [0.1, 0.15, 0.2]$ for the LLR method as suggested by Ren et al. [28] and report all these results. Ren et al. use another model-parameter λ which is the coefficient of L_2 regularization in training the LSTM model [28]. Ren et al. reported that λ is optional and does not markedly affect the overall performance of the LLR method compared to μ [28], we fix $\lambda = 0$ while changing μ to save computational resources. We also report the results based on $\lambda =$ 1e-4 which are very similar to the results based on $\lambda = 0$ for the viral and plasmid data sets in Supplementary Materials.

LLR was the state-of-the-art method and was the only available method that particularly focused on detecting OOD genomic sequences before MLR-OOD to the best of our knowledge. There are other methods developed for OOD tasks in vision that have not been adapted and evaluated in genomics. To make a comprehensive comparison with these methods, we compare MLR-OOD with three other methods that are commonly used for detecting OOD images. Specially, (1) Maximum of Softmax Probability (MSP)

based on a 1-dimensional convolutional neural networks (CNN) that classifies the in-distribution classes [22], (2) Deep Ensemble that takes the average of the predictive probabilities from 5 CNN classifiers trained based on different random initializations and random shuffling of training inputs [55], and (3) adjusted Out-of-DIstribution detector for Neural networks (ODIN) that uses temperature scaling and adds small perturbations to the input to enhance the separation of the softmax score distributions between ID and OOD samples [23]. This method was designed for continuous inputs and cannot be directly applied to discrete genomic sequences. We propose instead to add perturbations to the input of the last layer that is closest to the output of the neural network. We report the results in Supplementary Materials Section 2.2.

2.10 Evaluation metrics

We use two widely adopted metrics to evaluate the prediction accuracy for detecting OOD genomic sequences. The first one is the area under the receiver operating characteristic courve (**AUROC**). Assume we have calculated the predicting scores of whatever method for the testing sequences. For a given threshold we predict all the sequences having a prediction score higher than the threshold as ID sequences and otherwise as OOD sequences. The corresponding true positive rate (TPR) and false positive rate (FPR) are then calculated. By varying the threshold we can draw a receiver operating characteristic (ROC) curve with FPR and TPR as the two axes. Finally, we calculate the AUROC as the evaluation metric. The second one is the area under the precision-recall curve (**AUPRC**). The calculation of AUPRC is similar to that of AUROC except that we use precision and recall as the two axes. Higher values of both metrics indicate better model performance. Both metrics have been commonly used for evaluating the model performance of OOD detection [28, 22].

For each testing data set, we randomly select 1,000 ID and 1,000 OOD testing contigs from the testing data set and calculate the AUROC or AUPRC based on these 2,000 contigs. We repeat the process for 30 rounds and report the mean and standard deviation of both metrics.

3 Results

3.1 CNN classification models fail to distinguish between ID and OOD testing sequences

Although Ren et al. [28] have shown that deep neural network classifiers fail to properly deal with OOD bacterial sequences, we reproduce the conclusion and extend it to viral and plasmid sequences, showing that this phenomenon is universal in metagenomic sequence classification regardless of the sequence type. We use convolutionary neural network (CNN) as used by Ren et al. [28] to train classification models on the ID bacterial, viral, and plasmid data sets. We use the same CNN architecture as in Ren et al. [28]. Specifically, the CNNs contain one convolutional layer, one max-pooling layer, and a final dense layer with softmax activation for predicting class probabilities. The number of motifs for the convolutional layer is set as 1000 for the bacterial data set and 100 for the viral and plasmid data sets since there are much fewer ID training sequences in these two data sets. We monitor the ID validation loss that is calculated every 100 epoches up to 100,000 epoches and choose the epoch yielding the smallest validation loss. Since the viral and plasmid data sets lack an ID validation data set, we split each ID training data set and use 90% of the sequences for training the classification models and the remaining 10% for validation. Then we use the trained classification models to calculate the maximum softmax probability (MSP) $p(\tilde{y}|\mathbf{x}) = \max_k p(y = k|\mathbf{x}), 1 \leq k \leq K$ on the ID and OOD testing data sets. Conceptually ID testing sequences should have much higher MSP than OOD testing sequences if the classifiers work well.

Figure 2 shows the MSP of ID and OOD 250bp testing sequences based on the CNN classification models in the bacterial, plasmid, and viral data sets. As shown in the figure, ID sequences have slightly higher MSP than OOD sequences in the bacterial **Test2016** and **Test2018** testing data sets. On the other hand, no clear separation between the distributions of ID and OOD sequences can be observed. For the plasmid and viral data sets, on the contrary, OOD sequences even have relatively higher MSP compared to ID sequences, indicating that OOD sequences are more likely to be classifier into one of the training classes than ID sequences. This figure clearly demonstrates that the CNN models may fail to distinguish OOD sequences from ID sequences, meaning that OOD sequences are very likely to be misclassified by deep learning models

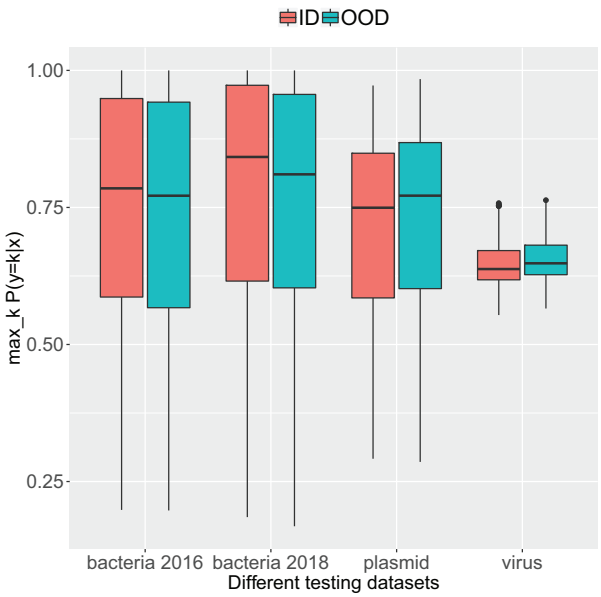

Figure 2: The bar plots of the maximum class probabilities of 10k ID and 10k OOD sequences in each type of testing data set. The contig length of the testing sequences is 250bp.

with high confidence. On the other hand, we also notice that the validation loss fluctuates markedly during training, meaning that the CNN classifiers are actually not stable for classifying new sequences. Therefore, OOD detection of genomic sequences is of great significance to microbial studies to ensure the credibility of taxonomic classification.

3.2 MLR-OOD outperforms LLR on all data sets

3.2.1 Markov order estimation for different testing data sets

The choice of Markov order for the MLR-OOD method depends on different testing data sets and contig lengths. For each testing data set and contig length, we use Bayesian Information Criterion (BIC) (details shown in Supplementary Materials) to estimate the Markov order of each contig and choose the most common estimated order for calculating $S^r_{\text{MLR-OOD}}(\mathbf{x})$ in equation (3). The estimated orders based on BIC have been shown to perform well in molecular sequence analyses [59].

For the **Test2016** bacterial data set, zero-th order (i.i.d.) is estimated as the most common order for contigs with length less or equal to 1,000bp. The longer 2,500bp and 5,000bp contigs are most likely to be estimated as first order Markov chains. The conclusion is the same for the **Test2018** bacterial data set and the plasmid data set except that i.i.d. is the most common order for the 2,500bp contigs. For viruses, zero-th order (i.i.d.) is the most common order regardless of the contig length. The distributions of the estimated orders for all data sets are shown in Supplementary Materials.

3.2.2 The optimal training length for MLR-OOD is 250bp

We first present the prediction accuracy of our MLR-OOD method for OOD genomic sequence detection on different training and testing lengths. The corresponding LSTM models are trained using ID training sequences of lengths 100bp, 250bp, and 500bp in the bacterial data set. Testing contigs of lengths 1,000bp, 2,500bp, and 5,000bp are used to evaluate the performances of LSTM models based on different training lengths. As shown in Figure 3, models trained using 250bp sequences consistently perform the best in terms of prediction accuracy. On the other hand, we observe that the computational time for training the LSTM models is proportional to the training sequence length. Models trained using 100bp sequences perform

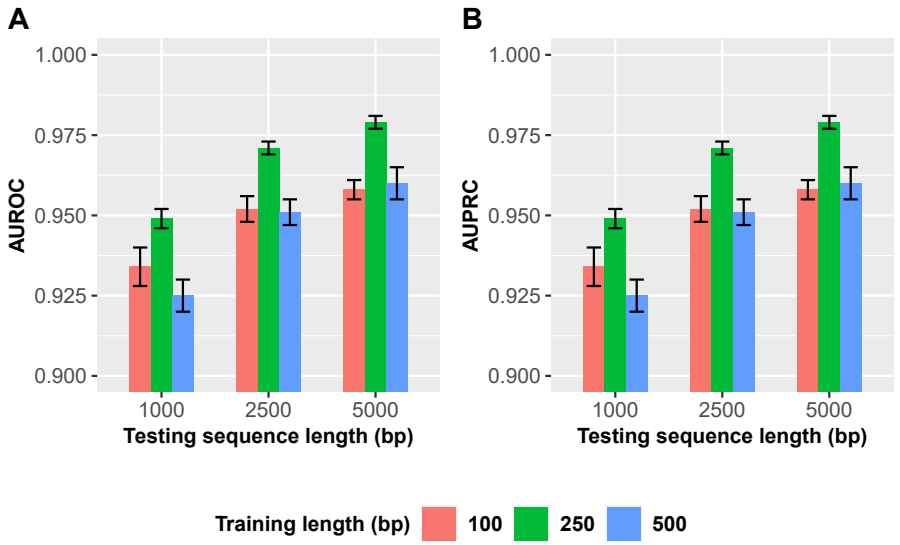

Figure 3: The prediction accuracy of MLR-OOD on the **Test2016** bacterial data set based on different training/testing sequence lengths. Each bar shows the mean accuracy of 30 random repetitions for a particular training/testing sequence length. Error bars indicate the standard deviation.

relatively lower with the least computational time. Models trained using 500bp sequences perform poorly in terms of both prediction accuracy and computational time. The poor performance of LSTM on long sequences (500bp) is probably due to the fact that the use of activation functions may result in gradient decay over layers [60]. Therefore, in this paper, we fix the training sequence length at 250bp as in Ren et al. [28] which gives excellent prediction accuracy and acceptable computational time.

3.2.3 The computational time of MLR-OOD

The computational time of MLR-OOD is dominated by training the generative models. Compared to training, the time for calculating the likelihoods of LSTM and Markov models is minimal. The computational time for training is proportional to both the number of epoches and the training sequence length. For example, training 900,000 epoches for lengths 100bp, 250bp, and 500bp sequences using the NVIDIA Tesla V100 GPU takes approximately 42, 105, and 210 hours, respectively. We recommend using GPU resources for training a large number of epoches in practice.

3.2.4 The comparison between MLR-OOD, max-LL, and LLR on different lengths of testing sequences

Next, we compare our MLR-OOD and max-LL methods with the LLR method using different lengths of testing contigs and present the results in Figure 4. The ID LSTM models are trained using the training data set of Ren et al. [28] and two corresponding testing data sets (**Test2016** and **Test2018**) are used for evaluating the performances of different methods. As shown in Figure 4 (A)-(D), our MLR-OOD method greatly outperforms the LLR method for detecting OOD genomic sequences, regardless of the contig length. Note that the max-LL method improves the prediction accuracy by a remarkable margin compared to the LLR method, revealing that considering the ID classes separately instead of using one general model for all ID classes is the major reason for the superiority of MLR-OOD. That being said, the Markov chain likelihood further increases the prediction accuracy as shown therein that MLR-OOD performs consistently better than max-LL, which can be possibly attributed to that Markov chain likelihoods reduced the effects of the confounding factors.

We would like to emphasize that MLR-OOD does not need to tune model-parameters and does not access

to the extra OOD validation data, while the LLR method does need to tune two extra model-parameters on the validation data set. For the comparison, we used the optimal model-parameters chosen by Ren et al. [28]. As for the contig length, all these three methods show an increasing trend of the prediction accuracy for longer contigs, which can be explained by that longer contigs generally contain more information for the particular taxons they belongs to. It is also worth mentioning that MLR-OOD yields very high prediction accuracy (AUROC and AUPRC > 0.9) when the testing contig length is at least 1,000bp, a length that many assembled metagenomic contigs can easily reach, clearly demonstrating that MLR-OOD is highly promising for detecting OOD bacterial sequences in real practice.

We present the results for predicting OOD viral sequences in Figure 4 (E) and (F). As shown therein, although the prediction accuracy still increases with the testing contig length, all these three methods generally yield a lower performance compared to the results of the bacterial data sets shown in Figure 4 (A)-(D). This can be possibly explained by the fact that viruses tend to have much shorter genomes but much higher mutation rates than their hosts [61, 62], making it more difficult to train LSTM models using limited amount of data for the ID classes to accurately capture the underlying taxonomic characteristics. Among these three methods, the two methods proposed in this paper: max-LL and MLR-OOD both perform much better than LLR in all scenarios, regardless of the choice of the model-parameter μ. Although the prediction accuracy of MLR-OOD is slightly lower than max-LL, the differences are minimal. This phenomenon can be explained by the fact that the prediction scores of max-LL of the viral contigs are already robust to the GC content, thus the adjustment by Markov likelihood is no longer needed, as we will show later in Figure 5. In general, max-LL and MLR-OOD are exchangeable for the viral data sets and both of them have noticeable improvement compared to the LLR method. Note that $\mu = 0.1$ performs better than $\mu = 0.15$ and $\mu = 0.2$ among the three different models of LLR.

Finally, we compare the model performances of the three methods on the plasmid data sets in Figure 4 (G) and (H). It is clearly shown in Figure 4 (G) and (H) that the prediction accuracy of these methods remains relatively low for short contigs compared to the bacterial data set, but increases rapidly if the contigs become longer than 1,000bp. Similar to the results of the viral data sets, we see that both max-LL and MLR-OOD achieve higher prediction accuracy than the LLR method even if compared with the model-parameters yielding the highest accuracy for LLR. Unlike the viral data sets, MLR-OOD slightly outperforms max-LL. The differences are almost negligible though. We will see in Figure 5 that both max-LL and MLR-OOD are robust to the GC content compared to LLR. For LLR, it is interesting to observe that $\mu = 0.2$ performs better than $\mu = 0.1$ and $\mu = 0.15$ for this data set, which is different from the viral data sets. This phenomenon indicates that the choice of the model-parameter μ using an extra validation data set is essential for LLR, as different data sets require different optimal choices of μ. In conclusion, MLR-OOD has better model performances than LLR even without using the information from extra validation data sets.

We also compare MLR-OOD with other state-of-the-art vision OOD detection methods adapted to genomic sequences including MSP, Deep Ensemble CNN, and adjusted ODIN. The results are shown in Table S3 in Supplementary Materials. Consistent with the results from Ren et al. [28], these methods do not perform as well as LLR and let alone MLR-OOD for bacterial sequences. For viruses and plasmids, the AUROC scores of the classifier-based methods are even lower than 0.5 as the mean score for ID sequences is even lower than that for OOD sequences as shown in Figure 2.

3.3 MLR-OOD is robust to the GC content on all data sets

In this section, we show that the MLR-OOD method is robust to the GC content that is a major confounding effect for detecting OOD genomic sequences. Figure 5 shows the relationship between the GC content and the prediction scores of LLR, max-LL and MLR-OOD defined in equations 1-3 on the bacterial, viral and plasmid testing data sets.

For the bacterial **Test2016** data set in which Ren et al. showed that the LLR method was robust to the GC content, we observe the same pattern they presented in [28] for LLR. As for max-LL, it is shown that the separation between ID and OOD sequences becomes clearer compared to LLR. However, the prediction score of max-LL is slightly biased by the GC content as OOD sequences having GC content between 0.4 and 0.7 tend to have slightly lower prediction scores. In contrast, our MLR-OOD method is less biased by the GC content compared to max-LL while maintaining a much better separation than LLR, explaining the best performance among the three methods shown in Figure 4 (A) and (B).

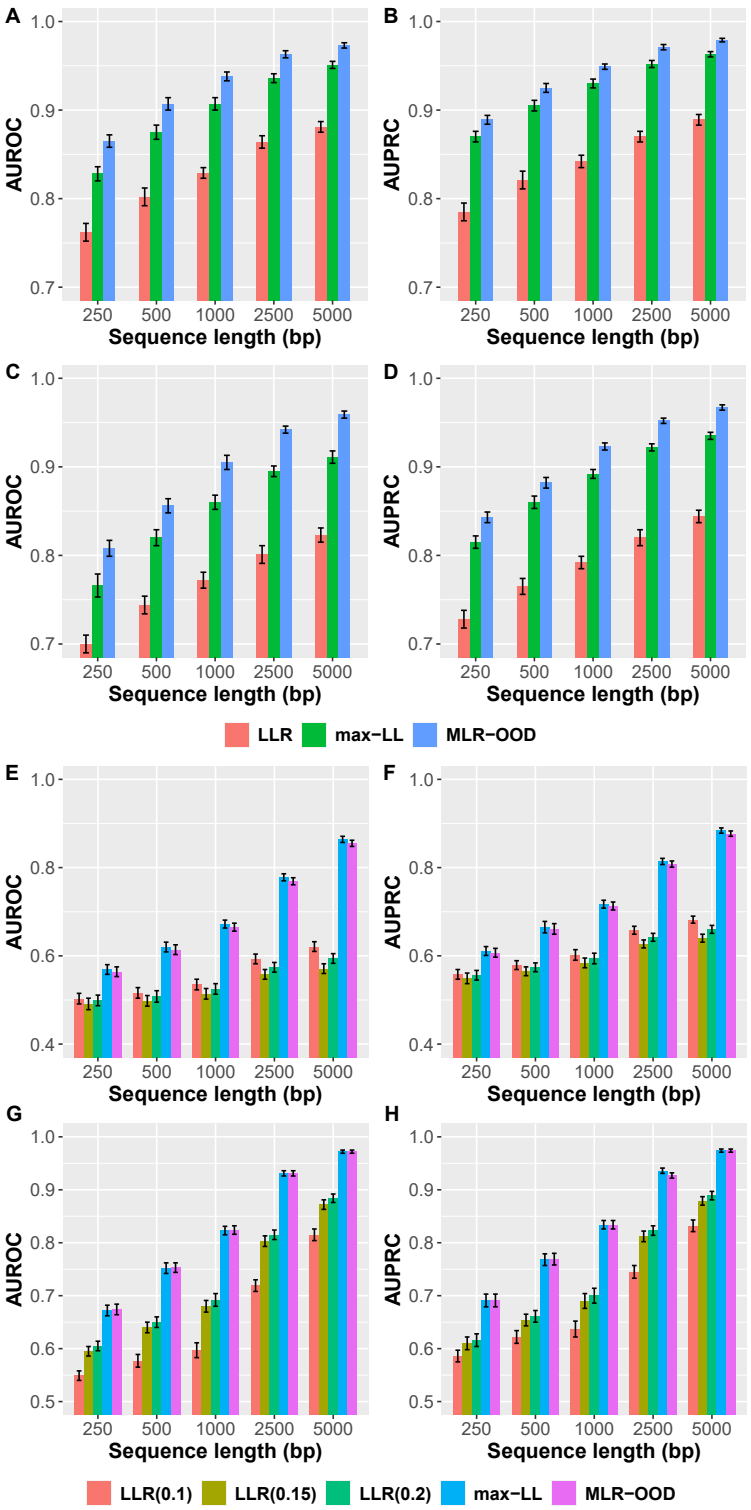

Figure 4: The prediction accuracies of LLR, max-LL, and MLR-OOD for OOD genomic sequences detection on the **Test2016**, **Test2018** bacterial data sets, viral datasets, and the plasmid datasets. (A) and (B): AUROC and AUPRC for the bacterial **Test2016** data set. (C) and (D): AUROC and AUPRC for the bacterial **Test2018** data set. (E) and (F): AUROC and AUPRC for the viral dataset. (G) and (H): AUROC and AUPRC for the plasmid data set. Each bar shows the mean accuracy of 30 random repetitions for each method and a particular sequence length. Error bars indicate the standard deviation.

The viral data sets and the plasmid testing data sets, nevertheless, display different patterns from the bacterial **Test2016** data set. First, the LLR method is biased by the GC content for both data sets. For the plasmid data set, there is a slightly increasing trend between the GC content and the prediction score. For the viral data set, the trend is reversed. Second, the prediction scores of max-LL of testing sequences in both data sets are not obviously associated with the GC content, possibly explaining the decent performances of max-LL on these two data sets. Third, as shown in Figure 5, the prediction scores of MLR-OOD are similar to those of max-LL, which is consistent to the results shown in Figure 4 that max-LL and MLR-OOD have very close prediction accuracy on the viral and plasmid data sets. This is understandable because the prediction score of max-LL is less biased by GC content compared to the bacterial data sets and adjustment using Markov chain likelihoods does not markedly improve the prediction scores. That being said, MLR-OOD still makes the prediction score slightly more independent of the GC content for the plasmid testing data set.

3.4 The prediction accuracy of MLR-OOD increases with the Mash distance threshold for choosing the OOD testing sequences

After constraining the OOD testing sequences to have minimum Mash distance to the ID classes greater than a certain threshold, we observe in Figure 6 that the prediction accuracy for OOD genomic sequences increases with the Mash distance threshold. For bacterial sequences which have been studied by Ren et al. based on the d_2^S distance [28], the trend is consistent as shown in Figure 6 (A) and (B). It is shown therein that the prediction accuracy increases from threshold 0 (no constrain) to 0.3 and then remains stable. For viral sequences shown in Figure 6 (C) and (D), the slightly increasing trend is similar. The plasmid sequences are shown in Figure 6 (E) and (F) to have the most obvious increment. The AUROC and AUPRC increase by more than 0.1 from Mash distance threshold 0 to 0.3 and then remain stable. We choose different cutoff thresholds for different types of sequences because that the distributions of the Mash distances for bacterial, viral, and plasmid sequences are different and we choose these cutoffs where the majority of OOD testing sequences have minimum Mash distance to ID classes. The histograms of the minimum Mash distance of OOD testing sequences to ID classes of bacterial, viral, and plasmid sequences are shown in Supplementary Materials Figure S1.

3.5 The impact of chimeric contigs on the prediction score and prediction accuracy of MLR-OOD

In this section, we show that the prediction score of MLR-OOD increases with the fraction of ID sequence fragments for chimeric contigs. As a result, the prediction accuracy of MLR-OOD also changes accordingly. Figure 7 illustrates the impact of chimeric contigs on the prediction score and prediction accuracy of MLR-OOD. The distributions of the prediction scores of MLR-OOD for bacterial, viral, and plasmid chimeric contigs are shown in Figure 7 (A), (C), and (E), respectively. It is shown that there is an obvious increasing trend with the fraction of ID sequence fragments for both bacterial and plasmid chimeric contigs. For viral chimeric contigs the trend is not as clear, which is possibly due to the fact that the prediction accuracy of MLR-OOD on the viral data set is generally low as shown in Figure 4. However, we still see generally more contigs receive higher prediction scores when the fraction of ID sequence fragments increases. This phenomenon is understandable since ID sequence fragments generally receive higher prediction scores than OOD sequence fragments. We also quantify the trend by calculating the AUROC for classifying completely OOD contigs (ID fraction $c = 0$) and contigs containing a certain fraction of ID sequences ($c = 0.25$, 0.5, 0.75, and 1) based on the prediction scores of MLR-OOD. As shown in Figure 7 (B) and (F), the AUROC increases remarkably from classifying completely OOD contigs and chimeric contigs with ID fraction 0.25 to classifying completely OOD and ID (ID fraction $c = 1$) contigs. For viral chimeric contigs shown in Figure 7 (D), the trend is also monotonically increasing although the slope is relatively low. These results are reasonable and consistent with the results shown in Figure 7 (A), (C), and (E). In conclusion, chimeric contigs receive intermediate prediction scores between completely ID and OOD contigs from MLR-OOD and the prediction accuracy depends on the fraction of ID/OOD sequences in those contigs.

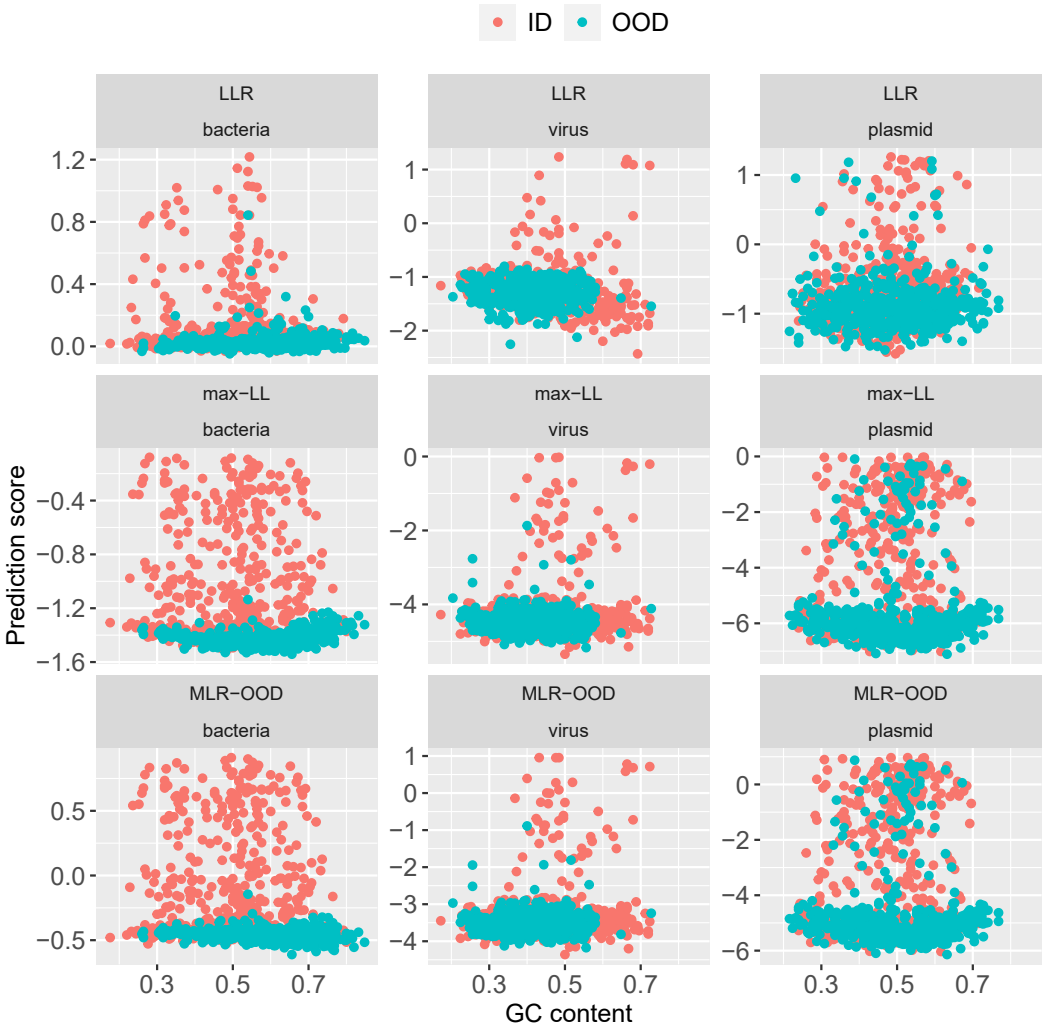

Figure 5: The relationship between the GC content and the prediction scores of LLR, max-LL and MLR-OOD on three types of testing sequences. Each subfigure contains 500 randomly selected ID and 500 randomly selected OOD 250bp testing sequences. For the bacterial data set, we use the **Test2016** data set. For the viral and the plasmid data sets, we select the prediction score corresponding to the μ yielding the highest prediction, that is, $\mu = 0.1$ for the viral data sets and $\mu = 0.2$ for the plasmid data sets.

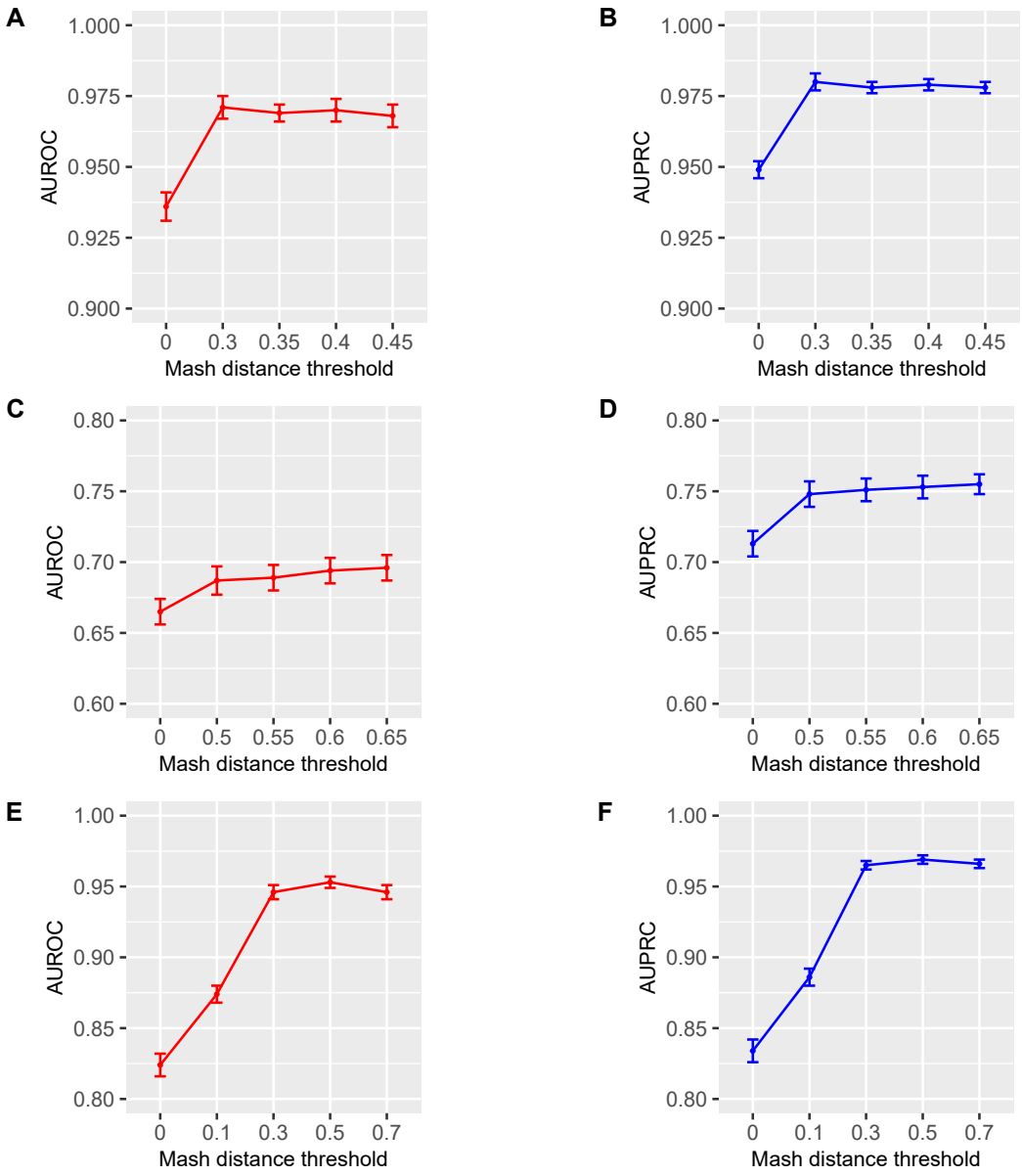

Figure 6: The relationship between the Mash distance threshold for choosing OOD sequences and the prediction accuracy of MLR-OOD on the bacterial **Test2016**, viral, and the plasmid datasets. (A) and (B): AUROC and AUPRC for the bacterial **Test2016** dataset. (C) and (D): AUROC and AUPRC for the viral dataset. (E) and (F): AUROC and AUPRC for the plasmid dataset. Each point shows the mean accuracy of 30 random repetitions for each method and a particular sequence length. Error bars indicate the standard deviation. The x-axis represents the minimum Mash distance threshold for choosing OOD testing genomes. The contig length is fixed as 1,000bp.

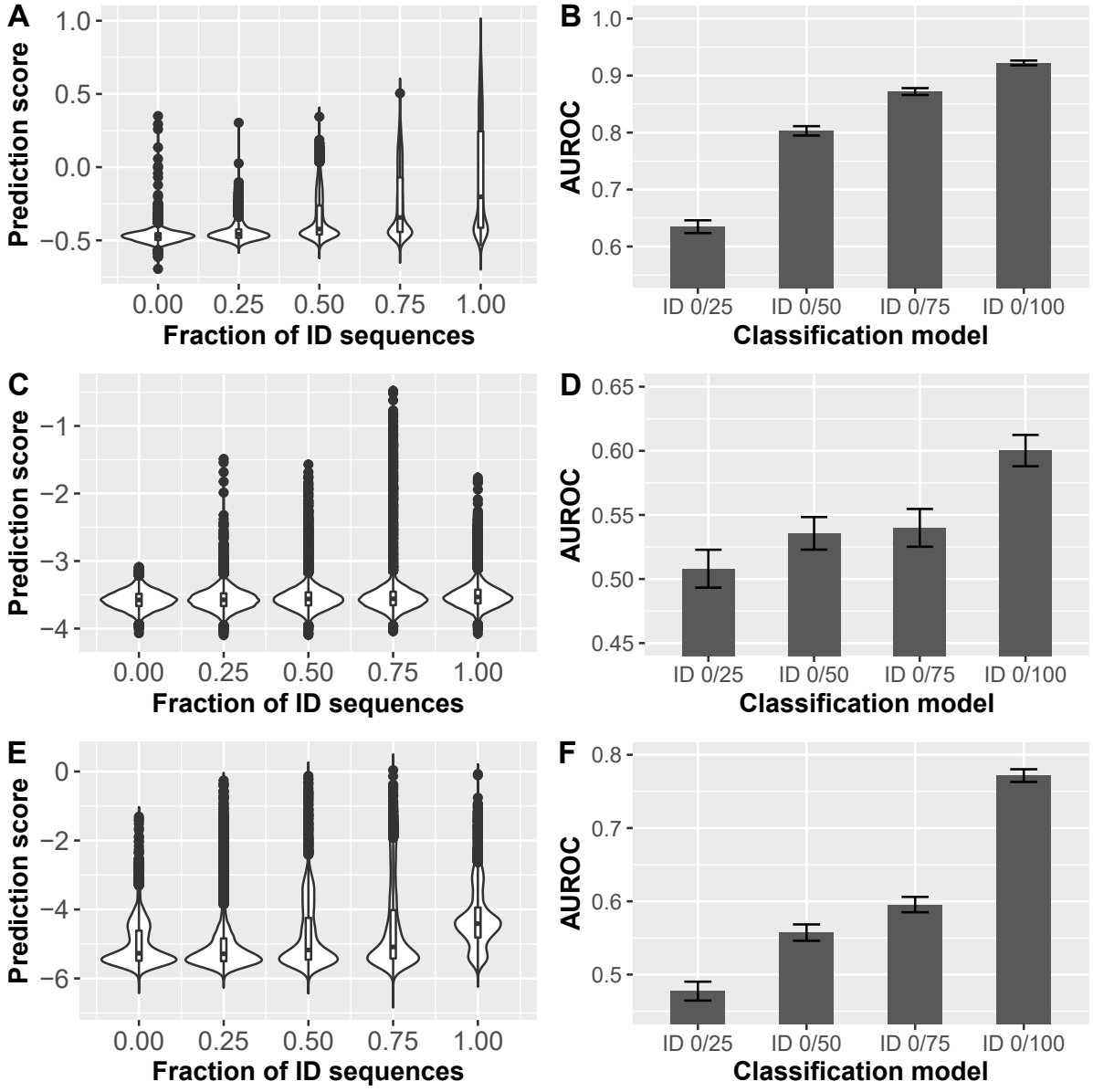

Figure 7: The impact of chimeric contigs on the prediction score and prediction accuracy of MLR-OOD. (A), (C), and (E): the violin plots of the prediction scores of bacterial **Test2016**, viral, and plasmid chimeric contigs, respectively. The x-axis represents the fraction of ID sequences in the chimeric contigs. The contig length is fixed at 1,000bp. (B), (D), and (F): the AUROC for classifying bacterial **Test2016**, viral, and plasmid contigs containing different fractions of ID sequences, respectively. For example, "ID 0/25" represents classifying contigs containing 0% and 25% ID sequences. Each bar shows the mean accuracy of 30 repetitions based on 1k randomly drawn testing contigs from each chimeric contig set. Error bars indicate the standard deviation.

Table 2: The comparison among MLR-OOD, max-LL, and LLR in several aspects showing their strengths and weaknesses.

Methods	MLR-OOD	max-LL	LLR
Accuracy	High	High	Relatively low
Parameter tuning	No	No	Two model-parameters to be tuned manually
Effect of GC content	Highly robust	Relatively robust	Not robust on the viral and plasmid datasets
Computational resource	High	High	Low

4 Discussion and Conclusions

Machine learning or deep learning models have been gaining in popularity for classifying microbial sequences because of their power in learning sequence patterns and generality to discover unknown sequences. However, their weakness in dealing with OOD sequences has been long neglected. Several studies show that deep learning classifiers are likely to classify OOD sequences into one of the training classes with high confidence, revealing that the detection of OOD genomic sequences is urgently needed. In this paper, we propose MLR-OOD, a Markov chain based likelihood ratio method to tackle this problem. We summarize the strengths and weaknesses of the three methods for detecting OOD genomic sequences: MLR-OOD and max-LL proposed by us and the LLR method proposed by Ren er al. [28], in different aspects in Table 2. Compared to the LLR method, the first work particularly addressing the detection of OOD genomic sequences, MLR-OOD has several key advantages. First, MLR-OOD utilizes the specific ID class labels by training a generative model for each ID class separately, making it possible to more precisely capture the distribution of ID sequences. Second, MLR-OOD bypasses tuning model-parameters by using the Markov chain likelihoods to adjust the likelihoods given by the ID models. This is of paramount importance for real applications since the assumption that part of the OOD data are accessible for tuning model-parameters is questionable in reality. Third, we show that the prediction score of MLR-OOD is robust to the GC content which is a main confounding effect for detecting OOD genomic sequences. Fourth, MLR-OOD consistently achieves remarkably higher prediction accuracy compared to the LLR method on all testing data sets even if LLR is based on the optimal model-parameters chosen from the validation data sets, clearly revealing the state-of-the-art performance of MLR-OOD. Compared to the max-LL method, the prediction accuracy gain of MLR-OOD is minimal on the viral and plasmid data sets though, possibly because the effect of adjustment based on sequence complexity is low on these data sets. Fifth, in addition to the bacterial data set composed by Ren et al. [28], we also construct a more updated bacterial testing data set along with the viral and plasmid data sets for comprehensively benchmarking current and future OOD detection methods.

Despite these key advantages, there are also some limitations for MLR-OOD. First, the training of the generative models for the ID classes requires a large amount of computational time and resource, especially when there are many ID classes. Second, just like other deep learning based methods, MLR-OOD needs a large training data set to avoid overfitting. For example, we observe that the prediction accuracy of MLR-OOD on the viral data set dropped compared to the bacterial data set, which can possibly be explained by the fact that the viral data set contains much fewer training sequences and thus overfitting may happen. We acknowledge that using overlapping sequences chopped from genomes rather than nonoverlapping sequences for training the generative models may be a better alternative in such scenarios. Third, it is difficult for MLR-OOD to increase the prediction accuracy if the maximum of the class conditional likelihoods is already robust to the GC content, just as shown on the plasmid and viral data sets. In the future, we hope to develop an updated version of MLR-OOD which can save the computational resource without compromising on the prediction accuracy. We also expect to combine novel methods targeting optimizing the model parameters of deep neural networks [63, 64, 65, 66] with MLR-OOD to make it more robust and powerful.

Code availability

The MLR-OOD software package is available at `https://github.com/xinbaiusc/MLR-OOD`.

Data availability

The metagenomic data consisting of bacterial, viral, and plasmid sequences for benchmarking the detection of OOD genomic sequences are available at
`https://drive.google.com/drive/folders/1KzOkQ_D1VWYqA-GDld783O7H8AzNuHkC?usp=sharing`.

Acknowledgements

We would like to thank Dr. Yingying Fan at the University of Southern California for helpful discussions. This research utilized GPU resources of Center for Advanced Research Computing (CARC), which is supported by the University of Southern California.

Funding

This research was partially supported by US National Institutes of Health (NIH) [R01GM120624, 1R01GM131407] and National Science Foundation (NSF) EF-2125142.

Competing interests

The authors declare no competing interests.

References

[1] Derrick E Wood and Steven L Salzberg. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biology, 15(3):R46, 2014.

[2] Derrick E Wood, Jennifer Lu, and Ben Langmead. Improved metagenomic analysis with Kraken 2. Genome Biology, 20(1):257, 2019.

[3] Daehwan Kim, Li Song, Florian P Breitwieser, and Steven L Salzberg. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Research, 26(12):1721-1729, 2016.

[4] Peter Menzel, Kim Lee Ng, and Anders Krogh. Fast and sensitive taxonomic classification for metagenomics with kaiju. Nature Communications, 7(1):11257, 2016.

[5] Florian P Breitwieser, DN Baker, and Steven L Salzberg. KrakenUniq: confident and fast metagenomics classification using unique k-mer counts. Genome Biology, 19(1):198, 2018.

[6] Rachid Ounit, Steve Wanamaker, Timothy J Close, and Stefano Lonardi. Clark: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genomics, 16(1):236, 2015.

[7] Robert J Robbins, Leonard Krishtalka, and John C Wooley. Advances in biodiversity: metagenomics and the unveiling of biological dark matter. Standards in Genomic Sciences, 11(1):69, 2016.

[8] Corie Lok. Mining the microbial dark matter. Nature News, 522(7556):270, 2015.

[9] Lindsey Solden, Karen Lloyd, and Kelly Wrighton. The bright side of microbial dark matter: lessons learned from the uncultivated majority. Current Opinion in Microbiology, 31:217-226, 2016.

[10] Zifan Zhu, Jie Ren, Sonia Michail, and Fengzhu Sun. MicroPro: using metagenomic unmapped reads to provide insights into human microbiota and disease associations. Genome Biology, 20(1):154, 2019.

[11] Paul B Eckburg, Elisabeth M Bik, Charles N Bernstein, Elizabeth Purdom, Les Dethlefsen, Michael Sargent, Steven R Gill, Karen E Nelson, and David A Relman. Diversity of the human intestinal microbial flora. Science, 308(5728):1635-1638, 2005.

[12] Stephen Nayfach, Zhou Jason Shi, Rekha Seshadri, Katherine S Pollard, and Nikos C Kyrpides. New insights from uncultivated genomes of the global human gut microbiome. Nature, 568(7753):505-510, 2019.

[13] Bas E Dutilh, Alejandro Reyes, Richard J Hall, and Katrine L Whiteson. Virus discovery by metagenomics: the (im) possibilities. Frontiers in Microbiology, 8:1710, 2017.

[14] Qiaoxing Liang, Paul W Bible, Yu Liu, Bin Zou, and Lai Wei. DeepMicrobes: taxonomic classification for metagenomics with deep learning. NAR Genomics and Bioinformatics, 2(1):lqaa009, 2020.

[15] Antonino Fiannaca, Laura La Paglia, Massimo La Rosa, Giovanni Renda, Riccardo Rizzo, Salvatore Gaglio, Alfonso Urso, et al. Deep learning models for bacteria taxonomic classification of metagenomic data. BMC bioinformatics, 19(Suppl 7):198, 2018.

[16] Jie Ren, Kai Song, Chao Deng, Nathan A Ahlgren, Jed A Fuhrman, Yi Li, Xiaohui Xie, Ryan Poplin, and Fengzhu Sun. Identifying viruses from metagenomic data using deep learning. Quantitative Biology, 8(1):64-77, 2020.

[17] Zhencheng Fang, Jie Tan, Shufang Wu, Mo Li, Congmin Xu, Zhongjie Xie, and Huaiqiu Zhu. PPR-Meta: a tool for identifying phages and plasmids from metagenomic fragments using deep learning. Gigascience, 8(6):giz066, 2019.

[18] Gregory Ditzler, Robi Polikar, and Gail Rosen. Multi-layer and recursive neural networks for metagenomic classification. IEEE Transactions on Nanobioscience, 14(6):608-616, 2015.

[19] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, 2014.

[20] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 427-436, 2015.

[21] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In International Conference on Machine Learning, pages 1321-1330. PMLR, 2017.

[22] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. arXiv preprint arXiv:1610.02136, 2016.

[23] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. International Conference on Learning Representations (ICLR), 2018.

[24] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. arXiv preprint arXiv:1807.03888, 2018.

[25] Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized ODIN: Detecting out-ofdistribution image without learning from out-of-distribution data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10951-10960, 2020.

[26] Apoorv Vyas, Nataraj Jammalamadaka, Xia Zhu, Dipankar Das, Bharat Kaul, and Theodore L Willke. Out-of-distribution detection using an ensemble of self supervised leave-out classifiers. In Proceedings of the European Conference on Computer Vision (ECCV), pages 550-564, 2018.

[27] Gabi Shalev, Yossi Adi, and Joseph Keshet. Out-of-distribution detection using multiple semantic label representations. arXiv preprint arXiv:1808.06664, 2018.

[28] Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. In Advances in Neural Information Processing Systems, pages 14680-14691, 2019.

[29] Terrance DeVries and Graham W Taylor. Learning confidence for out-of-distribution detection in neural networks. arXiv preprint arXiv:1802.04865, 2018.

[30] Sara Cuadros-Orellana, Laura Rabelo Leite, Ash Smith, Julliane Dutra Medeiros, Fernanda Badotti, Paula LC Fonseca, Aline BM Vaz, Guilherme Oliveira, and Aristóteles Góes-Neto. Assessment of fungal diversity in the environment using metagenomics: a decade in review. Fungal Genomics & Biology, 3(2):1, 2013.

[31] Paul D Donovan, Gabriel Gonzalez, Desmond G Higgins, Geraldine Butler, and Kimihito Ito. Identification of fungi in shotgun metagenomics datasets. PLoS One, 13(2):e0192898, 2018.

[32] Limin Fu, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics, 28(23):3150-3152, 2012.

[33] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. Neural Computation, 9(8):1735-1780, 1997.

[34] Neda Tavakoli. Modeling genome data using bidirectional LSTM. In 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC), volume 2, pages 183-188. IEEE, 2019.

[35] Dmitry Grapov, Johannes Fahrmann, Kwanjeera Wanichthanarak, and Sakda Khoomrung. Rise of deep learning for genomic, proteomic, and metabolomic data integration in precision medicine. Omics: a journal of integrative biology, 22(10):630-636, 2018.

[36] Jack Lanchantin, Ritambhara Singh, Beilun Wang, and Yanjun Qi. Deep motif dashboard: Visualizing and understanding genomic sequences using deep neural networks. In Pacific Symposium on Biocomputing 2017, pages 254-265. World Scientific, 2017.

[37] Joan Serrà, David Álvarez, Vicenç Gómez, Olga Slizovskaia, José F Núñez, and Jordi Luque. Input complexity and out-of-distribution detection with likelihood-based generative models. arXiv preprint arXiv:1909.11480, 2019.

[38] Olga G Troyanskaya, Ora Arbell, Yair Koren, Gad M Landau, and Alexander Bolshoy. Sequence complexity profiles of prokaryotic genomic sequences: A fast algorithm for calculating linguistic complexity. Bioinformatics, 18(5):679-688, 2002.

[39] Yuri L Orlov and Vladimir N Potapov. Complexity: an internet resource for analysis of DNA sequence complexity. Nucleic Acids Research, 32(suppl_2):W628-W633, 2004.

[40] Hagai Almagor. A Markov analysis of DNA sequences. Journal of Theoretical Biology, 104(4):633-645, 1983.

[41] Jonathan Arnold, A Jamie Cuticchia, David A Newsome, W Wesley Jennings III, and Robert Ivarie. Mono-through hexanucleotide composition of the sense strand of yeast DNA: a Markov chain analysis. Nucleic Acids Research, 16(14):7145-7158, 1988.

[42] PJ Avery. The analysis of intron data and their use in the detection of short signals. Journal of Molecular Evolution, 26(4):335-340, 1987.

[43] Peter J Avery and Daniel A Henderson. Fitting Markov chain models to discrete state series such as DNA sequences. Journal of the Royal Statistical Society: Series C (Applied Statistics), 48(1):53-61, 1999.

[44] B Edwin Blaisdell. A measure of the similarity of sets of sequences not requiring sequence alignment. Proceedings of the National Academy of Sciences, 83(14):5155-5159, 1986.

[45] B Edwin Blaisdell. Markov chain analysis finds a significant influence of neighboring bases on the occurrence of a base in eucaryotic nuclear DNA sequences both protein-coding and noncoding. Journal of Molecular Evolution, 21(3):278-288, 1985.

[46] Gesine Reinert, Sophie Schbath, and Michael S Waterman. Probabilistic and statistical properties of words: an overview. Journal of Computational Biology, 7(1-2):1-46, 2000.

[47] Michael S Waterman. Introduction to computational biology: maps, sequences and genomes. CRC Press, 1995.

[48] RichardWKatz. On some criteria for estimating the order of a Markov chain. Technometrics, 23(3):243-249, 1981.

[49] Huaiqiu Zhu, Qian Guo, Mo Li, Chunhui Wang, Zhengcheng Fang, Peihong Wang, Jie Tan, Shufang Wu, and Yonghong Xiao. Host and infectivity prediction of Wuhan 2019 novel coronavirus using deep learning algorithm. BioRxiv, 2020.

[50] Gesine Reinert, David Chew, Fengzhu Sun, and Michael S Waterman. Alignment-free sequence comparison (i): statistics and power. Journal of Computational Biology, 16(12):1615-1634, 2009.

[51] Jie Ren, Xin Bai, Yang Young Lu, Kujin Tang, Ying Wang, Gesine Reinert, and Fengzhu Sun. Alignment-free sequence analysis and applications. Annual Review of Biomedical Data Science, 1:93-114, 2018.

[52] Brian D Ondov, Todd J Treangen, Páll Melsted, Adam B Mallonee, Nicholas H Bergman, Sergey Koren, and Adam M Phillippy. Mash: fast genome and metagenome distance estimation using minhash. Genome Biology, 17(1):132, 2016.

[53] Martin Ayling, Matthew D Clark, and Richard M Leggett. New approaches for metagenome assembly with short reads. Briefings in Bioinformatics, 21(2):584-594, 2020.

[54] Joshua A Udall and R Kelly Dawe. Is it ordered correctly? validating genome assemblies by optical mapping. The Plant Cell, 30(1):7-14, 2018.

[55] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in Neural Information Processing Systems 30, Pages 6405–6416, (2017).

[56] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. arXiv preprint arXiv:1711.09325, 2017.

[57] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. arXiv preprint arXiv:1812.04606, 2018.

[58] Hyunsun Choi, Eric Jang, and Alexander A Alemi. Waic, but why? generative ensembles for robust anomaly detection. arXiv preprint arXiv:1810.01392, 2018.

[59] Leelavati Narlikar, Nidhi Mehta, Sanjeev Galande, and Mihir Arjunwadkar. One size does not fit all: on how Markov model order dictates performance of genomic sequence analyses. Nucleic Acids Research, 41(3):1416-1424, 2013.

[60] Shuai Li, Wanqing Li, Chris Cook, Ce Zhu, and Yanbo Gao. Independently recurrent neural network (IndRNN): Building a longer and deeper rnn. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5457-5466, 2018.

[61] Siobain Duffy. Why are RNA virus mutation rates so damn high? PLoS Biology, 16(8):e3000003, 2018.

[62] Kayla M Peck and Adam S Lauring. Complexities of viral mutation rates. Journal of Virology, 92(14), 2018.

[63] Steven R Young, Derek C Rose, Thomas P Karnowski, Seung-Hwan Lim, and Robert M Patton. Optimizing deep learning hyper-parameters through an evolutionary algorithm. In Proceedings of the Workshop on Machine Learning in High-Performance Computing Environments, pages 1-5, 2015.

[64] Ilija Ilievski, Taimoor Akhtar, Jiashi Feng, and Christine Shoemaker. Efficient hyperparameter optimization for deep learning algorithms using deterministic rbf surrogates. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 31, 2017.

[65] Dougal Maclaurin, David Duvenaud, and Ryan Adams. Gradient-based hyperparameter optimization through reversible learning. In International Conference on Machine Learning, pages 2113-2122. PMLR, 2015.

[66] Tobias Domhan, Jost Tobias Springenberg, and Frank Hutter. Speeding up automatic hyperparameter optimization of deep neural networks by extrapolation of learning curves. In Twenty-fourth International Joint Conference on Artificial Intelligence, 2015.