RESEARCH ARTICLE

qiRNApredictor: A Novel Computational Program for the Prediction of qiRNAs in *Neurospora crassa*

Haiyou Deng, Quan Liu, Wei Cao, Rong Gui, Chengzhang Ma, Ming Yi, Yuangen Yao*

Department of Physics, College of Science, Huazhong Agricultural University, Wuhan, Hubei, China

* yyg@mail.hzau.edu.cn

Citation: Deng H, Liu Q, Cao W, Gui R, Ma C, Yi M, et al. (2016) qiRNApredictor: A Novel Computational Program for the Prediction of qiRNAs in *Neurospora crassa*. PLoS ONE 11(7): e0159487. doi:10.1371/journal.pone.0159487

Editor: Vinod Scaria, CSIR Institute of Genomics and Integrative Biology, INDIA

Received: June 12, 2015

Accepted: July 5, 2016

Published: July 18, 2016

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by grants from Innovation Fund of Huazhong Agricultural University (2662014BQ069 YY; 2662015BQ045 HD). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Recently, a new type of small interfering RNAs (qiRNAs) of typically 20~21 nucleotides was found in *Neurospora crassa* and rice and has been shown to regulate gene silencing in the DNA damage response. Identification of qiRNAs is fundamental for dissecting regulatory functions and molecular mechanisms. In contrast to other expensive and time-consuming experimental methods, the computational prediction of qiRNAs is a conveniently rapid method for gaining valuable information for a subsequent experimental verification. However, no tool existed to date for the prediction of qiRNAs. To this purpose, we developed the novel qiRNA prediction software package qiRNApredictor. This software demonstrates a promising sensitivity of 93.55% and a specificity of 71.61% from the leave-one-out validation. These studies might be beneficial for further experimental investigation. Furthermore, the local package of qiRNApredictor was implemented and made freely available to the academic community at Supplementary material.

Introduction

Small non-coding RNAs (sRNAs) of 20~30 nucleotides (nt) have gained significant attention in recent years as they are widely involved in various biological processes such as the embryonic, neuronal, muscle, and germline development [1–3]. QDE-2-interacting small RNAs (qiRNAs) of typically 20~21 nt are a new class of sRNAs. qiRNAs are induced by DNA damage [4] and mediate gene silencing in the DNA damage response (DDR) pathway in *Neurospora crassa* by inhibiting protein translation [4]. Although qiRNAs were first discovered in 2009, only little is known about their biogenesis and functionality. It has been demonstrated that the biogenesis of qiRNAs requires DNA-damage-induced aberrant RNAs (aRNAs) as precursors [4]. Moreover, RNA-dependent RNA polymerase QDE-1, the Werner and Bloom RecQ DNA helicase homologue QDE-3, as well as dicers have been previously shown to be involved in the production of qiRNAs [4]. After DNA damage, aRNAs are highly induced and then specifically recognized by RNA-dependent polymerases to produce double-stranded RNAs (dsRNAs). Subsequently, dsRNAs are converted to qiRNAs by dicers [4, 5]. However, the detailed mechanism for the generation of qiRNAs is largely unknown and requires elucidation in further experimental studies.

Previously, novel qiRNAs in *Neurospora crassa* [4] or rice were identified almost exclusively by immunoprecipitation followed by sequencing. However, this conventional approach for the identification of qiRNAs is time-consuming, labor-intensive, and inefficient and even holds the risk of not detecting lowly expressed or issue-specific qiRNAs. Computational methods can overcome these experimental hurdles and extract general information from known qiRNAs to predict novel qiRNAs for further experimental manipulation [6]. However, as far as we know, no qiRNA prediction tool has been implemented so far. Therefore, the development of efficient computational approaches for qiRNA prediction is urgently needed and intriguing.

Common sequence or structure conservation-based approaches in microRNAs (miRNAs) prediction cannot be directly adopted for qiRNA prediction as no evidence has been reported for a high evolutionary conservation of the sequence and structure of qiRNAs across species. Unlike Piwi-interacting RNAs (piRNAs), the density of plot of qiRNAs does not show a striking clustering characteristic [4]. Thus, no clustering characteristic can be used for qiRNA prediction. However, it has been demonstrated that qiRNAs exhibit strong position-specific preferences for uracil (U) at the first nucleotide of the 5' end and for adenine (A) at the first nucleotide of the 3' end [4]. In bioinformatics, position-specific nucleotide preferences are usually employed for sRNAs prediction. For example, ping-pong-dependent piRNAs show some position-specific nucleotide preferences, such as for thymine (T) at the first position (1T) and for A at the 10th position (10A) [6]. To capture and leverage these sequence features for piRNA prediction, Betel *et al.* constructed a 21×4 feature vector based on a 21-base window around the 5' end (plus 10 nt upstream and 10 nt downstream) and trained a support vector machine (SVM) model for piRNA prediction [7]. In 2011, Zhang *et al.* used a simple *k*-mer scheme to construct 1364 dimension feature vectors for describing the candidate piRNA sequences and then constructed the software piRNApredictor for piRNA prediction based on an improved Fisher linear algorithm [6]. Moreover, position-specific scoring matrix (PSSM) is a commonly used representation of sequence motifs, which has been widely applied to the prediction of significant biological signals such as DNA binding sites [8], RNA binding sites [9], promoters [10], protein secondary structure prediction [11], etc.

In this work, we developed an ingenious method for the extraction of features based on position probability matrices (PPM) of biosequences and then constructed a novel qiRNApredictor (qiRNA predictor) software package with 80 features. Firstly, the experimentally verified qiRNAs were collected manually from the scientific literatures and then used for training the Random forest (RF) model using the randomForest R package. Subsequently, the performance and robustness of qiRNApredictor were extensively evaluated by *n*-fold cross-validations (CV) as well as the leave-one-out cross-validation (LOO-CV). Upon LOO-CV, qiRNApredictor exhibits a promising sensitivity of 93.55% and a specificity of 71.61%, which can be used for the prediction of qiRNAs.

Materials and Methods

Data preparation

Herein, only qiRNA sequences in *Neurospora crassa* were considered (no qiRNA in rice was sequenced in the work by Chen *et al.* [5]). Lee *et al.* identified 184 individual qiRNA sequences in *Neurospora crassa* [4]. However, 23 qiRNA sequences were directly discarded as they included undefined bases. We further eliminated six individual qiRNA sequences as they were identical to other qiRNA sequences. Finally, 155 experimentally verified qiRNAs were collected as positive samples to train a RF model. Three negative datasets were built. For the convenience of discussion, we called them the "Random", the "sRNA-segment", and "milRNA" negative datasets, according to the ways we collected them or the data origin. The "Random" negative dataset was randomly extracted from NONCODE database. As previously described [6], non-coding RNA sequences obtained from the NONCODE database (version 3.0) [12] were fragmented into non-overlapping segments. For each of these non-qiRNA segments, we shuffled it 10000 times to destroy any potentially functional structures [6]. Then, the same amount of negative samples were randomly selected from the segments under the constraint condition that the length distribution of the selected segments was identical with that of positive qiRNAs. The "sRNA-segement" negative dataset was collected from Rfam database. The sequences of sRNAs of *Neurospora crassa* were retrieved from Rfam database (release 12.0) [13]. Because the lengths of sRNAs are much longer than that of qiRNAs, we fragmented the sequences of sRNAs into non-overlapping segments under the constraint that the length distribution of sRNA segments was similar to that of positive qiRNAs. Then, the same amount of sRNA segments were randomly selected and regarded as the analogue of the degraded fragments of sRNAs. miRNA-like small RNAs (milRNA) are about 19~25 nt and have a strong preference (51.08%) for U at the 5' end. They were firstly discovered in *Neurospora crassa* and called milRNAs due to their similarities with miRNAs [14]. Furthermore, 325 milRNAs were manually collected from the scientific literatures to construct the "milRNA" negative dataset. Since only 25 milRNAs were obtained from *Neurospora crassa*, we also collected the putative milRNAs from other fungi species, including *Trichoderma reesei* [15], *Sclerotinia sclerotiorum* [16], *Metarhizium anisopliae* [17], *Zymoseptoria tritici* [18], *Fusarium oxysporum* [19], *Fusarium graminearum* [20], *Antrodia cinnamomea* [21], *Aspergillus flavus* [22], *Penicillium chrysogenum* [23], and *Penicillium marneffei* [24]. Finally, these three negative datasets were incorporated with the positive dataset respectively to construct the "Random", "sRNA-segment" and "milRNA" training datasets.

Feature extraction

The extraction of an appropriate set of features for training a prediction model is one of the vital, yet most challenging issues in machine learning-based prediction approaches. To capture the position-specific preference of nucleotides, the occurrence probabilities of each nucleotide at the first ten positions (1 through 10) and the last ten positions (-1 to -10) of both positive and negative samples were calculated to create four PPMs (Fig 1). It should be noted that qiRNA sequences of less than 20 nt exhibit some overlaps of the first and last ten positions. The score of nucleotide *i* at position *j* can be calculated according to the formula

$$\mathrm{S}(\boldsymbol{i},\boldsymbol{j}) = \log_2\left(\frac{\boldsymbol{P}_{ij}}{\boldsymbol{N}_{ij}}\right),\quad \boldsymbol{i} \in [\mathrm{A},\mathrm{C},\mathrm{G},\mathrm{U}],\ \boldsymbol{j} \in [-10,\cdots,-1;\ 1,\cdots,10]$$

where $\boldsymbol{P}_{ij}$ and $\boldsymbol{N}_{ij}$ are the probability of nucleotide *i* at position *j* in PPMs produced by positive and negative samples, respectively (Fig 1). The score gives an indication how much the position-specific preference of positive samples differs from that of negative samples. For the cases with $\boldsymbol{P}_{ij} = \boldsymbol{N}_{ij}$, the score equals zero, which means this feature cannot offer any information for the prediction. However, when $\boldsymbol{P}_{ij}$ or $\boldsymbol{N}_{ij}$ is zero, it doesn't mean this feature has great significance. Instead, it may be simply because of the limited sample size. To circumvent the infinite value caused by inadequate sampling, we assigned a zero score for the cases that $\boldsymbol{P}_{ij}$ or $\boldsymbol{N}_{ij}$ is zero. In total, 80 scoring values were directly regarded as features to train a RF model. Then *F*-score was used to measure the discriminatory power of each feature above [25]. The *F*-score of

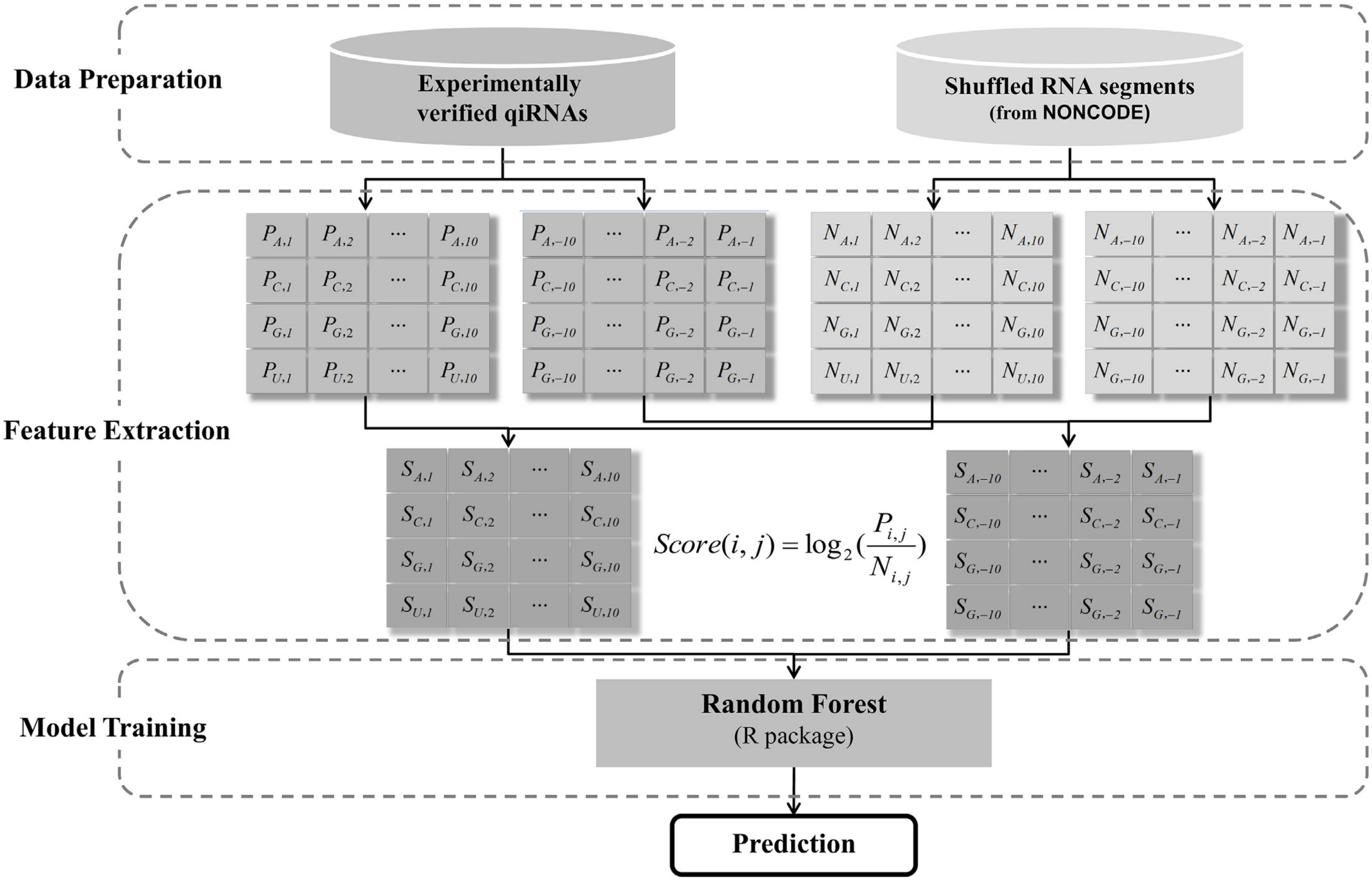
Data Preparation
Experimentally verified qiRNAs
Shuffled RNA segments (from NONCODE)
Feature Extraction
$P_{A,1}$ $P_{A,2}$ ··· $P_{A,10}$
$P_{C,1}$ $P_{C,2}$ ··· $P_{C,10}$
$P_{G,1}$ $P_{G,2}$ ··· $P_{G,10}$
$P_{U,1}$ $P_{U,2}$ ··· $P_{U,10}$
$P_{A,-10}$ ··· $P_{A,-2}$ $P_{A,-1}$
$P_{C,-10}$ ··· $P_{C,-2}$ $P_{C,-1}$
$P_{G,-10}$ ··· $P_{G,-2}$ $P_{G,-1}$
$P_{G,-10}$ ··· $P_{G,-2}$ $P_{G,-1}$
$N_{A,1}$ $N_{A,2}$ ··· $N_{A,10}$
$N_{C,1}$ $N_{C,2}$ ··· $N_{C,10}$
$N_{G,1}$ $N_{G,2}$ ··· $N_{G,10}$
$N_{U,1}$ $N_{U,2}$ ··· $N_{U,10}$
$N_{A,-10}$ ··· $N_{A,-2}$ $N_{A,-1}$
$N_{C,-10}$ ··· $N_{C,-2}$ $N_{C,-1}$
$N_{G,-10}$ ··· $N_{G,-2}$ $N_{G,-1}$
$N_{G,-10}$ ··· $N_{G,-2}$ $N_{G,-1}$
$S_{A,1}$ $S_{A,2}$ ··· $S_{A,10}$
$S_{C,1}$ $S_{C,2}$ ··· $S_{C,10}$
$S_{G,1}$ $S_{G,2}$ ··· $S_{G,10}$
$S_{U,1}$ $S_{U,2}$ ··· $S_{U,10}$
$Score(i,j) = \log_2(\frac{P_{i,j}}{N_{i,j}})$
$S_{A,-10}$ ··· $S_{A,-2}$ $S_{A,-1}$
$S_{C,-10}$ ··· $S_{C,-2}$ $S_{C,-1}$
$S_{G,-10}$ ··· $S_{G,-2}$ $S_{G,-1}$
$S_{G,-10}$ ··· $S_{G,-2}$ $S_{G,-1}$
Model Training
Random Forest (R package)
Prediction

Fig 1. The flowchart of qiRNApredictor.

doi:10.1371/journal.pone.0159487.g001

the *i*th feature is defined as:

$$F(i) \equiv \frac{(\bar{x}_i^{(+)} - \bar{x}_i)^2 + (\bar{x}_i^{(-)} - \bar{x}_i)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}(x_{k,i}^{(+)} - \bar{x}_i^{(+)})^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}(x_{k,i}^{(-)} - \bar{x}_i^{(-)})^2}$$

where n_+ and n_- are the numbers of positive and negative samples, respectively; $\bar{x}_i$, $\bar{x}_i^{(+)}$, and $\bar{x}_i^{(-)}$ are the average of the *i*th feature of total, positive, and negative samples, respectively. $x_{k,i}^{(+)}$ and $x_{k,i}^{(-)}$ are the *i*th feature of the *k*th positive and negative sample, respectively. Larger *F*-scores indicate better discrimination [25].

Random forest

RF is an ensemble learning classifier consisting of a multitude of tree-structured classifiers [26]. RF makes use of two powerful machine-learning methods—the bagging and random feature selection in the tree induction [27]. In the bagging algorithm, each tree is trained by using a bootstrap sample of the training data, and the prediction results of the ensemble are aggregated by majority vote or averaging rule to give the final prediction results [27]. Instead of using all features, only a random subset of features was used here to split at each node when growing a single tree. A type of CV together with the training step using out-of bag (OOB) samples was used to assess the prediction performance of RF. More specifically, a particular bootstrap sample was adopted to grow each tree during the process of training. Since bootstrapping sampling is sampling with replacement, some of the samples were ignored, while others were reused. The ignored samples constitute the OOB sample. On average, $1-e^{-1} \cong 2/3$ of the training samples was used for growing the tree leaving $e^{-1} \cong 1/3$ as OOB, which have not been used for tree construction. Therefore, these samples could be used to evaluate the prediction performance [26, 27]. The RF algorithm was implemented by the randomForest R package.

Performance evaluation

As previously described [28], among the predicted positive results obtained by qiRNApredictor, the real positives are called true positives (*TP*), while the other positive results are called false positives (*FP*). Among the predicted negative results obtained by qiRNApredictor, real negatives are called true negatives (*TN*), while the other negative results are called false negatives (*FN*). The performance is evaluated based on four measurements of specificity (*Sp*), sensitivity (*Sn*), accuracy (*Ac*), and Matthew's correlation coefficient (*MCC*). These indexes are defined as

$$Sn = \frac{TP}{TP + FN},\ Sp = \frac{TN}{TN + FP},\ Ac = \frac{TP + TN}{TP + FP + TN + FN}$$

and

$$MCC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}}$$

In this study, we performed 4-, 6-, 8-, and 10-fold CVs as well as the LOO-CV. Receiver Operating Characteristic (ROC) curves were plotted for performance visualization.

Results

A novel algorithm for qiRNA prediction

In this work, we collected experimentally identified qiRNAs from the scientific literatures. After eliminating redundancy in sequences and directly removing sequences that include undefined bases, a dataset of 155 experimentally verified qiRNA sequences was obtained (S1 Table). As previously described, a negative dataset containing 155 samples was constructed. Finally, a balanced database containing 155 positive and negative samples was used for model construction.

qiRNAs are very short of approximately 20~21 nt in length (Fig 2), which renders it very difficult to predict qiRNAs with high accuracy. However, the first and last nucleotide of qiRNAs have a high preference for U and A, respectively (Fig 3). Therefore, we attempted to predict qiRNAs with the position-specific preferences of nucleotides in qiRNA sequences. To this goal, PPMs were firstly constructed for characterizing position-specific properties of qiRNAs. Based on the PPMs, 80 log-likelihood scores were calculated as features for training prediction model (see details in Materials and Methods). Based on the training datasets, we used the *F*-score [25] to rank 80 features. As expected, nucleotides at the first and last position, such as 1U, 1A, 1C, or -1A, exhibit high *F*-scores (Fig 4). This is consistent with the position-specific preferences of nucleotides in qiRNA sequences, which demonstrates that we have succeeded in capturing these characteristics in qiRNA sequences.

To evaluate the performance and robustness of the qiRNApredictor, LOO-CV and 4-, 6-, 8-, and 10-fold CVs were performed. The LOO-CV results based on "Random" training dataset show that our method predicts at 93.55% sensitivity, 71.61% specificity, 82.58% accuracy and

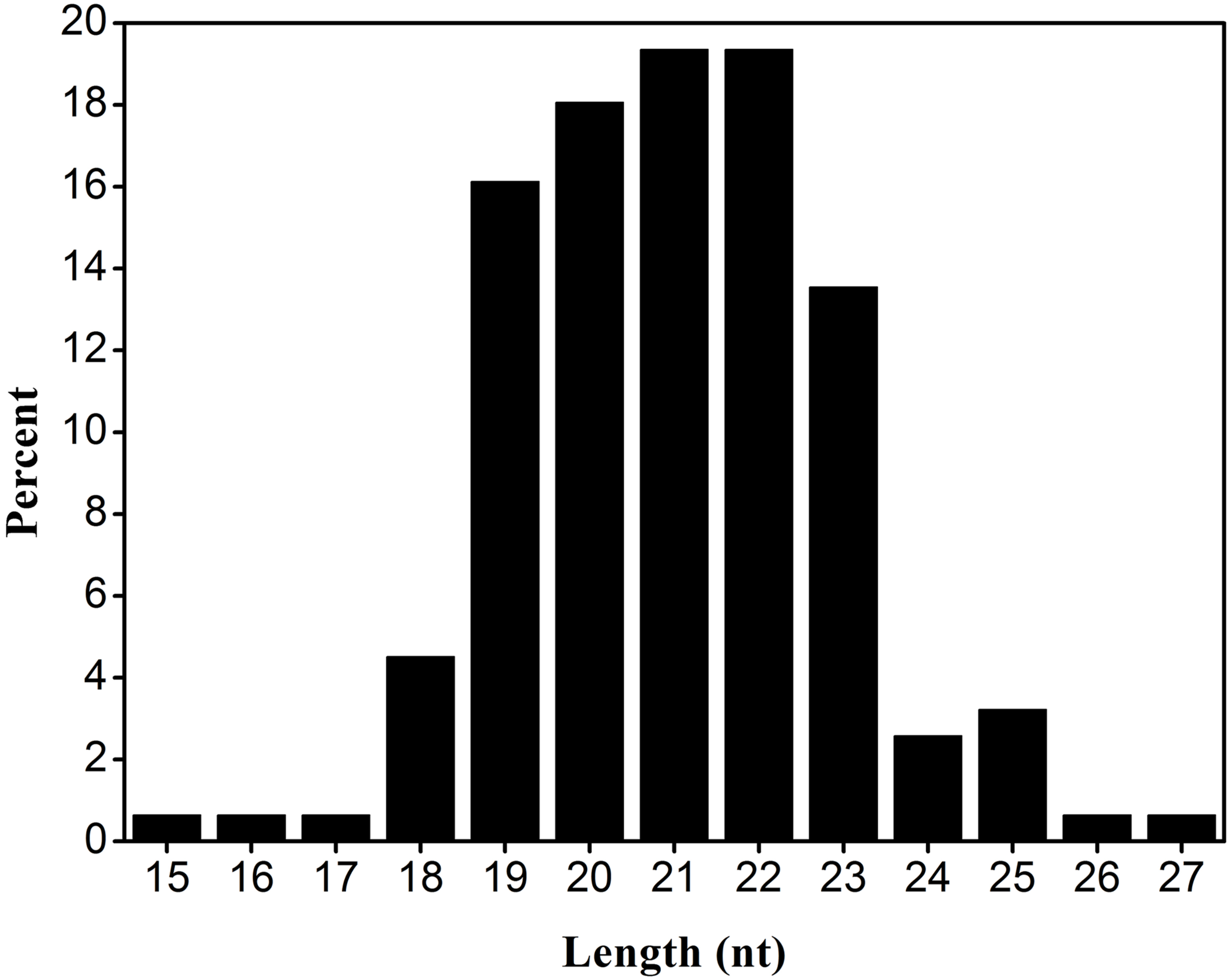

Fig 2. The length distribution of qiRNAs in *Neurospora crassa*.

doi:10.1371/journal.pone.0159487.g002

0.6679 *MCC* value (Table 1). The results of the 4-, 6-, 8-, and 10-fold CVs are also close to those of the LOO-CV. From the ROC curves (Fig 5A), AUC (area under ROC curves) values were calculated as 0.8779 (LOO-CV), 0.8772 (4-fold CV), 0.8752 (6-fold CV), 0.8761 (8-fold CV), and 0.8765 (10-fold CV), respectively (Fig 5A). As no qiRNA prediction tool existed to date, the performance of our qiRNA prediction tool could not be compared to any existing tool. To further test the ability of qiRNApredictor in distinguishing qiRNAs among other sRNAs, we also run it through "sRNA-segement" and "milRNA" datasets. milRNA is 19~25 nt in length with a strong preference (51.08%) for U at the 5' end, which is somewhat similar to qiRNA. As shown in Fig 5B & Table 1, the results demonstrate that the qiRNApredictor is able to identify qiRNAs among other sRNAs with quite significant AUC, *MCC*, sensitivity and specificity. Taken together, qiRNApredictor is a promising tool for the prediction of candidate qiRNAs.

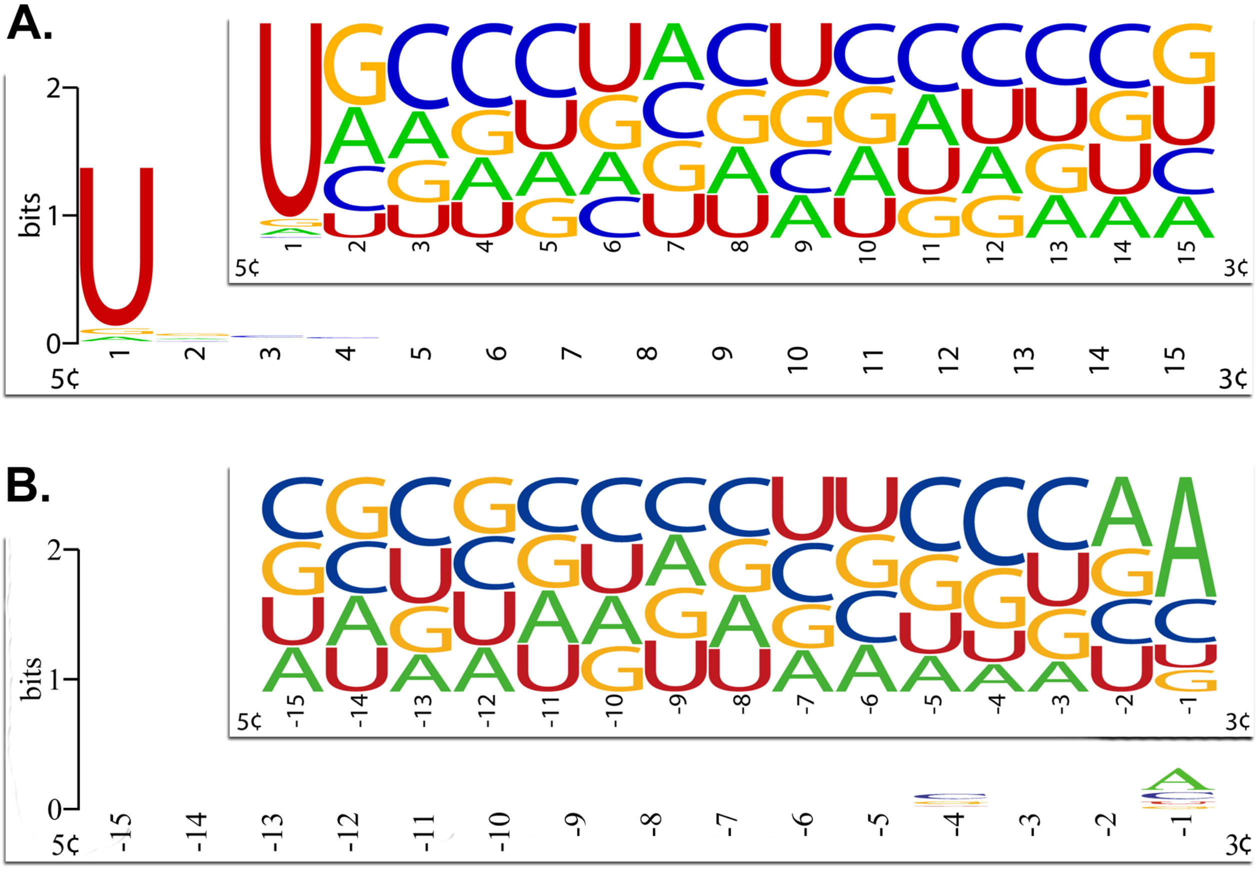

Fig 3. Position-specific nucleotide preferences of qiRNAs in *Neurospora crassa* for (A) the first 15 nt substring from 5' end to 3' end and (B) the last 15 nt substring from 5' end to 3' end. The insets are Frequency Plot of qiRNA sequences. The sequence logo analysis was implemented by WebLogo [29].

doi:10.1371/journal.pone.0159487.g003

Comparison to k-mer feature classes

In bioinformatics, *k*-mers usually refer to *k*-gram or *k*-tuples of DNA or protein sequences and can be used to find certain regions within biosequences or be employed as *k*-mer statistics for giving discrete probability distributions of possible *k*-mer combinations [6]. *K*-mers can be used to distinguish qiRNA from non-qiRNA based on differences of string usages between the different sequence classes. Here, 1–5 nt strings were used to characterize positive or negative sequences by a vector consisting of the frequencies of *k*-mer strings. We first adjusted the *k* parameter in *k*-mer feature classes to obtain a better performance based on the same training dataset. With increasing *k* values, the AUC values exhibit first a rapid increase, which slows down in the following (Fig 6A). Although the increase of features may result in over-fitting, the best performance (AUC = 0.7790) of *k*-mer feature classes is lower than that of PPM features classes (AUC = 0.8779) (Fig 6B). This demonstrates that PPM is more suitable for qiRNA prediction. Moreover, we combined PPM and *k*-mer feature classes for the prediction of qiRNAs. However, the prediction performance of the combined feature classes exhibits to be very close to that of single PPM feature classes (Fig 6B).

Discussion

As a new regulatory factor for mediating gene silencing in the DNA damage response, qiRNAs have increasingly attracted considerable interest and investigative efforts. The identification of

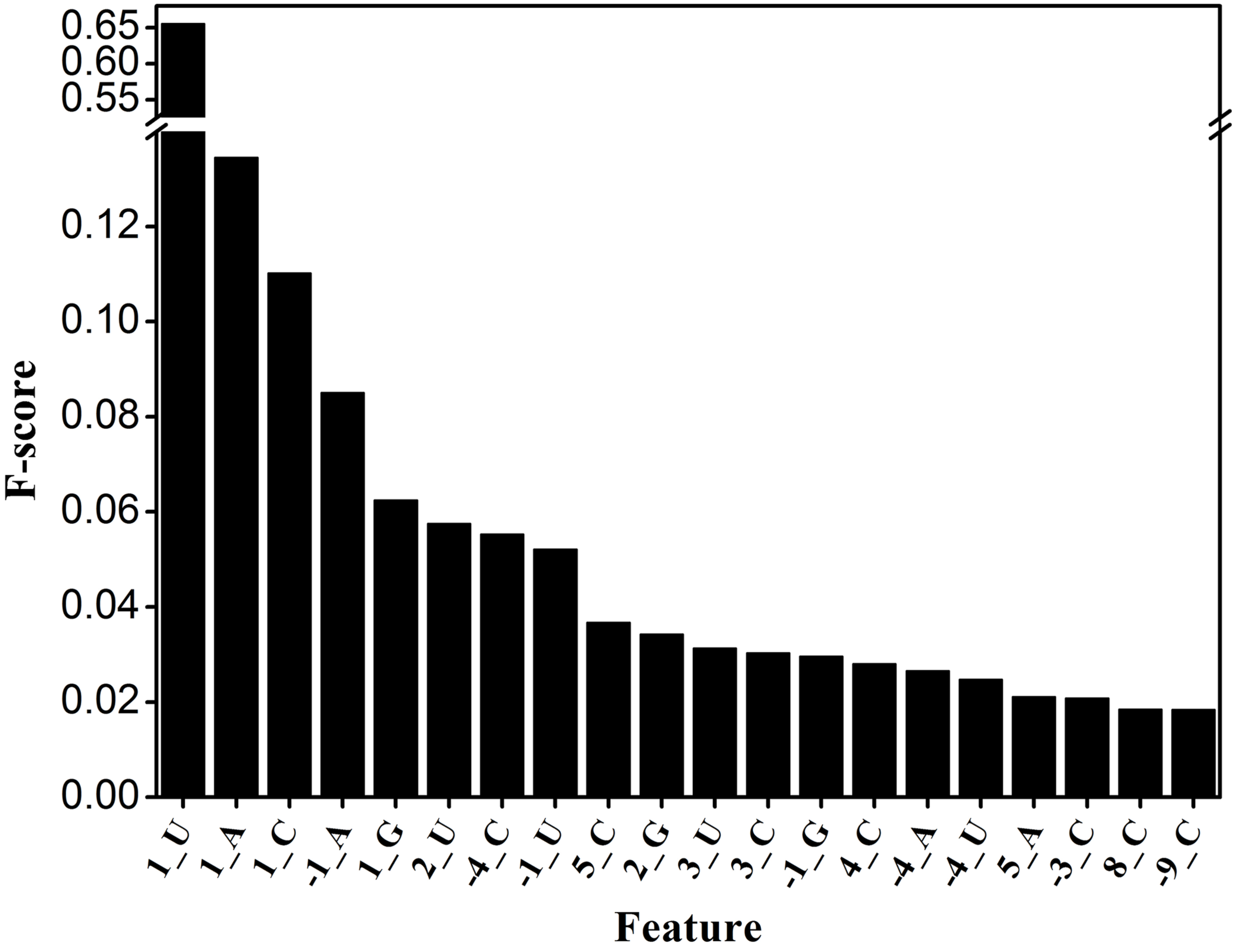

Fig 4. The *F*-scores of PPM features. Larger *F*-scores indicate better discrimination.

doi:10.1371/journal.pone.0159487.g004

qiRNAs is fundamental for dissecting regulatory functions and molecular mechanisms. Compared to other expensive and time-consuming experimental methods, the computational prediction of qiRNAs is a conveniently rapid method of getting useful information for subsequent experimental verification. However, no qiRNA prediction tool existed to date. Therefore, developing a novel approach for qiRNA prediction is required and very intriguing. In this work, we designed a novel software named qiRNApredictor based on PPM features and RF

Table 1. The performances of qiRNApredictor on each training dataset by the LOO CV.

Dataset	*Ac* (%)	*Sn* (%)	*Sp* (%)	*MCC*
"Random" training dataset	82.58	93.55	71.61	0.6679
"sRNA-segement" training dataset	80.97	86.45	75.48	0.6231
"milRNA" training dataset	79.17	70.32	83.38	0.5303

doi:10.1371/journal.pone.0159487.t001

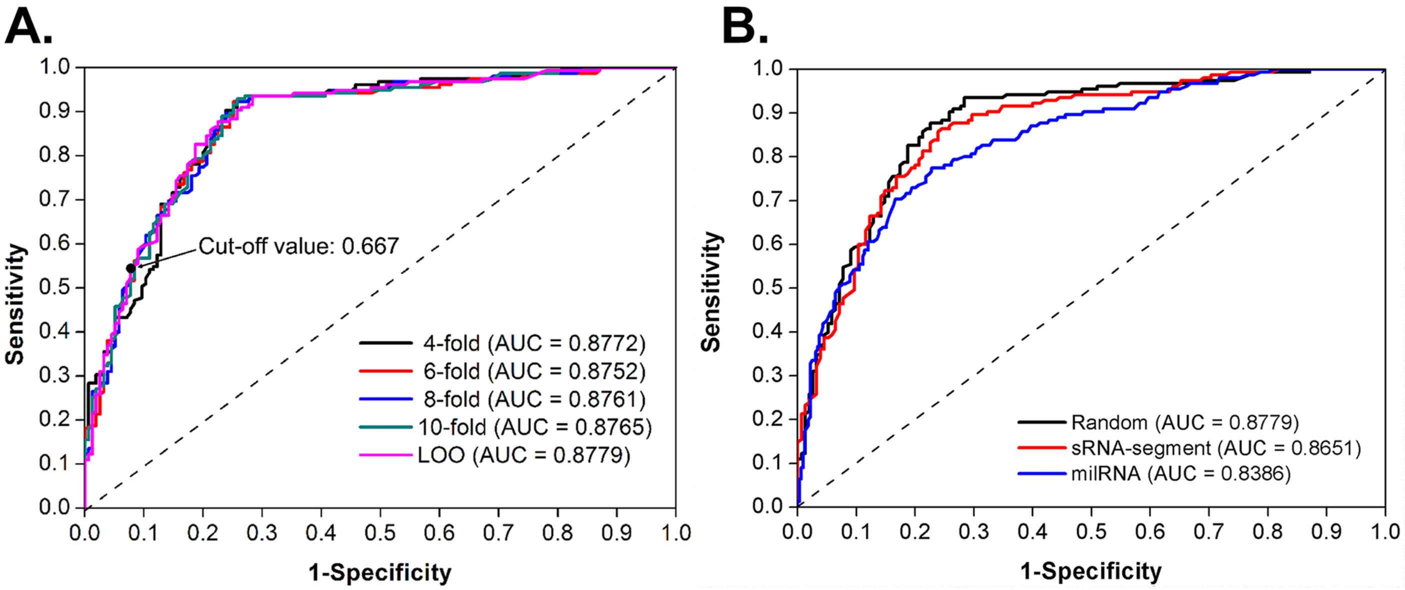

Fig 5. The prediction performance of qiRNApredictor. (A) The LOO-CV as well as 4-, 6-, 8-, and 10-fold CVs based on "Random" training dataset were calculated. The ROC curves and AUCs were also drawn and analyzed. To increase the specificity, we recommended a stringent cut-off value 0.667 for experimental investigation. The specificity of qiRNApredictor with the cut-off value 0.667 is 92.26%, while the sensitivity is 54.84%. (B) The ROC curves and AUCs were drawn and analyzed for the "Random", "sRNA-segment", and "milRNA" training datasets, respectively, by LOO-CV.

doi:10.1371/journal.pone.0159487.g005

algorithm. The performance and robustness of qiRNApredictor were extensively evaluated by n-fold CVs as well as LOO-CV, which gave very promising results.

Here, the window size refers to the length of the region of interest at both ends of the sequences. To evaluate the effect of window size on the prediction performance, the window size was adjusted, and the AUC value was calculated based on the results of 5-fold CV. The

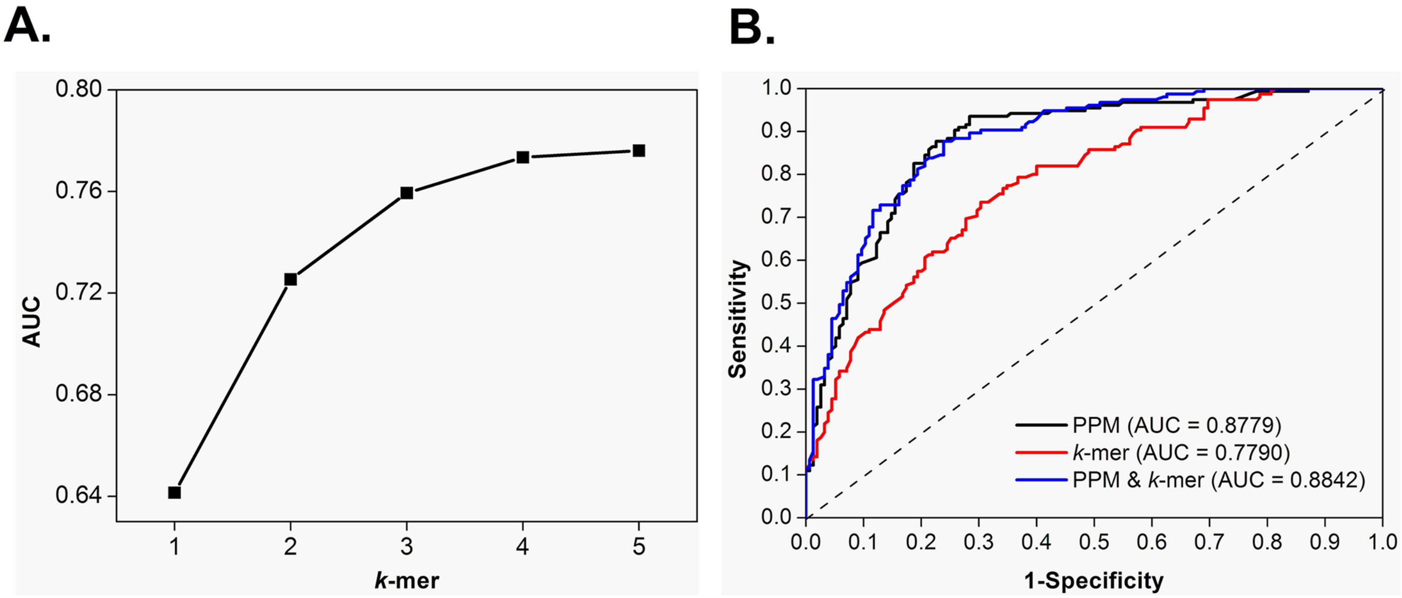

Fig 6. Comparison to *k*-mer feature classes. (A) Effect of the *k* parameter in *k*-mer feature classes on the prediction performance by 5-fold CV; (B) The ROC curves for RFs trained with different feature classes.

doi:10.1371/journal.pone.0159487.g006

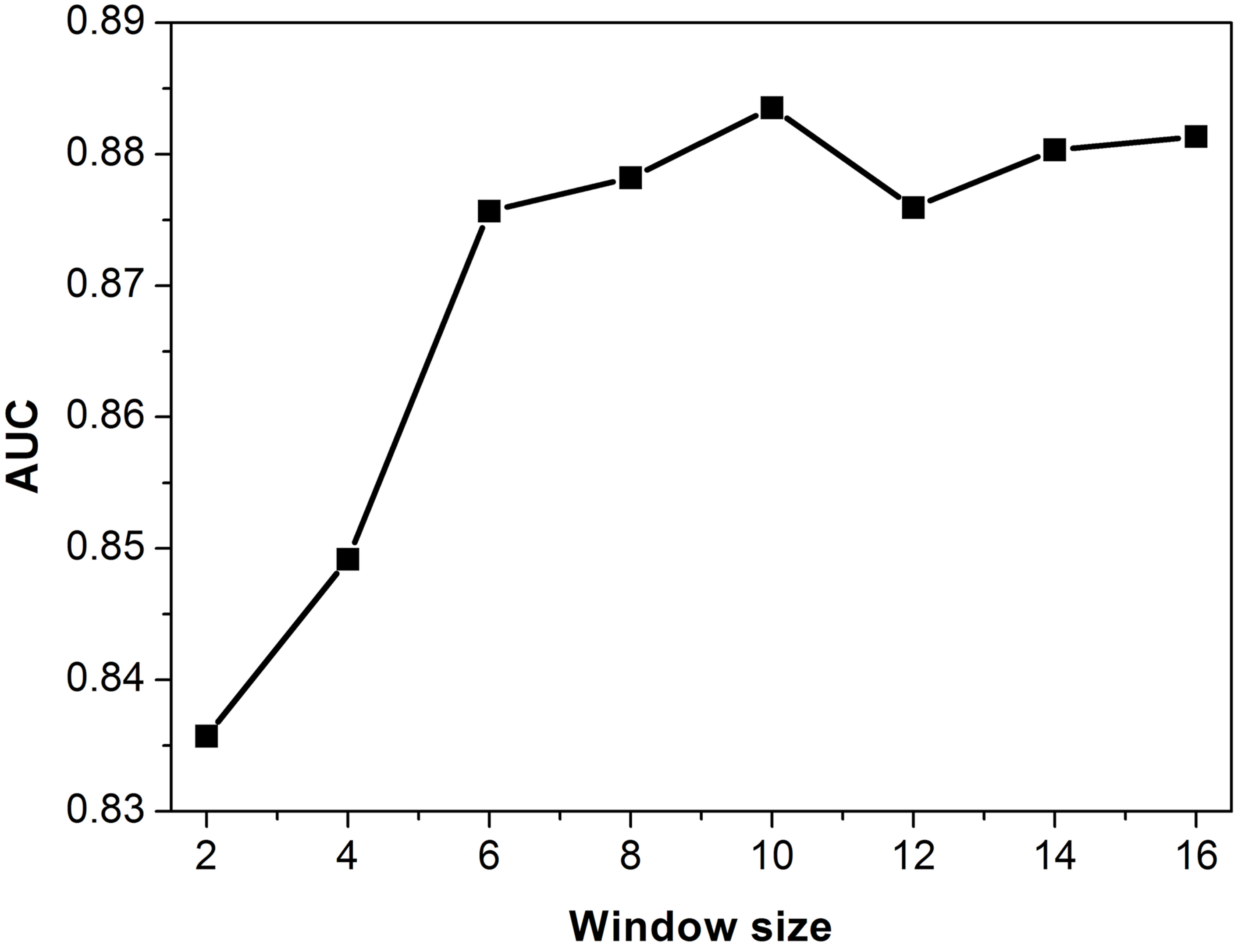

Fig 7. Effect of the window size in PPM feature classes on the prediction performance by 5-fold CV.

doi:10.1371/journal.pone.0159487.g007

AUC values first rapidly increase and then slow down with an increasing window size (Fig 7). There is a qiRNA sequence of only 15 nt, which was simply ignored when testing the window of 16 nt. Interestingly, the combination of two adjacent bases at both ends of sequences obtained a high AUC value of 0.8357. This either demonstrates that the characteristic signal at both ends of the qiRNA sequences is very significant, or it suggests that non-qiRNAs are not very similar to genuine qiRNAs. However, the approach of non-qiRNAs construction is standard according to previous studies [6].

The performance of a classifier system would be improved by selecting the most discriminative set of features due to reducing the complexity of the classifier system. In contrast, the simple combination of PPM and *k*-mer feature classes does not improve the performance of the classifier system. This suggests that the feature set exhibits redundancies, which leaves room for optimization of the feature set used in qiRNApredictor in the subsequent work. Furthermore, the prediction performance of qiRNApredictor would be improved by finding the secondary structure features of qiRNA precursors. Taken together, our studies provide a novel and promising approach for qiRNA prediction and will facilitate further functional studies of qiRNAs.

Supporting Information

S1 File. The compressed files in ZIP format of the local package of qiRNApredictor.
(ZIP)

S1 Table. The dataset of 155 experimentally verified qiRNAs in *Neurospora crassa* obtained from the work by Lee *at al* [4].
(XLS)

Acknowledgments

The authors thank Dr. Ya Jia (CCNU) and Dr. Anbang Li (CCNU) for their helpful suggestions on the project implementation.

Author Contributions

Conceived and designed the experiments: YY HD QL. Performed the experiments: YY HD. Analyzed the data: YY HD QL MY. Contributed reagents/materials/analysis tools: YY HD QL WC RG CM MY. Wrote the paper: YY HD.

References

1. Moazed D. Small RNAs in transcriptional gene silencing and genome defence. Nature. 2009; 457 (7228):413–20. Epub 2009/01/23. doi: 10.1038/nature07756 PMID: 19158787; PubMed Central PMCID: PMC3246369.
2. Stefani G, Slack FJ. Small non-coding RNAs in animal development. Nat Rev Mol Cell Biol. 2008; 9 (3):219–30. Epub 2008/02/14. doi: 10.1038/nrm2347 PMID: 18270516.
3. Pauli A, Rinn JL, Schier AF. Non-coding RNAs as regulators of embryogenesis. Nat Rev Genet. 2011; 12(2):136–49. Epub 2011/01/20. doi: 10.1038/nrg2904 PMID: 21245830; PubMed Central PMCID: PMC4081495.
4. Lee HC, Chang SS, Choudhary S, Aalto AP, Maiti M, Bamford DH, et al. qiRNA is a new type of small interfering RNA induced by DNA damage. Nature. 2009; 459(7244):274–7. Epub 2009/05/16. doi: 10.1038/nature08041 PMID: 19444217; PubMed Central PMCID: PMC2859615.
5. Chen H, Kobayashi K, Miyao A, Hirochika H, Yamaoka N, Nishiguchi M. Both OsRecQ1 and OsRDR1 are required for the production of small RNA in response to DNA-damage in rice. PloS one. 2013; 8(1): e55252. Epub 2013/02/06. doi: 10.1371/journal.pone.0055252 PMID: 23383126; PubMed Central PMCID: PMC3559376.
6. Zhang Y, Wang X, Kang L. A k-mer scheme to predict piRNAs and characterize locust piRNAs. Bioinformatics. 2011; 27(6):771–6. Epub 2011/01/13. doi: 10.1093/bioinformatics/btr016 PMID: 21224287; PubMed Central PMCID: PMC3051322.
7. Betel D, Sheridan R, Marks DS, Sander C. Computational analysis of mouse piRNA sequence and biogenesis. PLoS Comput Biol. 2007; 3(11):e222. Epub 2007/11/14. doi: 10.1371/journal.pcbi.0030222 PMID: 17997596; PubMed Central PMCID: PMC2065894.
8. Ahmad S, Sarai A. PSSM-based prediction of DNA binding sites in proteins. BMC Bioinformatics. 2005; 6:33. Epub 2005/02/22. doi: 10.1186/1471-2105-6-33 PMID: 15720719; PubMed Central PMCID: PMC550660.
9. Kumar M, Gromiha MM, Raghava GP. Prediction of RNA binding sites in a protein using SVM and PSSM profile. Proteins. 2008; 71(1):189–94. Epub 2007/10/13. doi: 10.1002/prot.21677 PMID: 17932917.
10. Hertzberg L, Zuk O, Getz G, Domany E. Finding motifs in promoter regions. J Comput Biol. 2005; 12 (3):314–30. Epub 2005/04/29. doi: 10.1089/cmb.2005.12.314 PMID: 15857245.
11. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol. 1999; 292(2):195–202. Epub 1999/09/24. doi: 10.1006/jmbi.1999.3091 PMID: 10493868.

12. Bu D, Yu K, Sun S, Xie C, Skogerbo G, Miao R, et al. NONCODE v3.0: integrative annotation of long noncoding RNAs. Nucleic acids research. 2012; 40(Database issue):D210–5. Epub 2011/12/03. doi: 10.1093/nar/gkr1175 PMID: 22135294; PubMed Central PMCID: PMC3245065.

13. Burge SW, Daub J, Eberhardt R, Tate J, Barquist L, Nawrocki EP, et al. Rfam 11.0: 10 years of RNA families. Nucleic acids research. 2013; 41(Database issue):D226–32. Epub 2012/11/06. doi: 10.1093/nar/gks1005 PMID: 23125362; PubMed Central PMCID: PMC3531072.

14. Lee HC, Li L, Gu W, Xue Z, Crosthwaite SK, Pertsemlidis A, et al. Diverse pathways generate microRNA-like RNAs and Dicer-independent small interfering RNAs in fungi. Molecular cell. 2010; 38 (6):803–14. Epub 2010/04/27. doi: 10.1016/j.molcel.2010.04.005 PMID: 20417140; PubMed Central PMCID: PMC2902691.

15. Kang K, Zhong J, Jiang L, Liu G, Gou CY, Wu Q, et al. Identification of microRNA-Like RNAs in the filamentous fungus Trichoderma reesei by solexa sequencing. PloS one. 2013; 8(10):e76288. Epub 2013/10/08. doi: 10.1371/journal.pone.0076288 PMID: 24098464; PubMed Central PMCID: PMC3788729.

16. Zhou J, Fu Y, Xie J, Li B, Jiang D, Li G, et al. Identification of microRNA-like RNAs in a plant pathogenic fungus Sclerotinia sclerotiorum by high-throughput sequencing. Molecular genetics and genomics: MGG. 2012; 287(4):275–82. Epub 2012/02/09. doi: 10.1007/s00438-012-0678-8 PMID: 22314800.

17. Zhou Q, Wang Z, Zhang J, Meng H, Huang B. Genome-wide identification and profiling of microRNA-like RNAs from Metarhizium anisopliae during development. Fungal biology. 2012; 116(11):1156–62. Epub 2012/11/17. doi: 10.1016/j.funbio.2012.09.001 PMID: 23153806.

18. Yang F. Genome-wide analysis of small RNAs in the wheat pathogenic fungus Zymoseptoria tritici. Fungal biology. 2015; 119(7):631–40. Epub 2015/06/11. doi: 10.1016/j.funbio.2015.03.008 PMID: 26058538.

19. Chen R, Jiang N, Jiang Q, Sun X, Wang Y, Zhang H, et al. Exploring microRNA-like small RNAs in the filamentous fungus Fusarium oxysporum. PloS one. 2014; 9(8):e104956. Epub 2014/08/21. doi: 10.1371/journal.pone.0104956 PMID: 25141304; PubMed Central PMCID: PMC4139310.

20. Chen Y, Gao Q, Huang M, Liu Y, Liu Z, Liu X, et al. Characterization of RNA silencing components in the plant pathogenic fungus Fusarium graminearum. Scientific reports. 2015; 5:12500. Epub 2015/07/28. doi: 10.1038/srep12500 PMID: 26212591; PubMed Central PMCID: PMC4515635.

21. Lin YL, Ma LT, Lee YR, Lin SS, Wang SY, Chang TT, et al. MicroRNA-like small RNAs prediction in the development of Antrodia cinnamomea. PloS one. 2015; 10(4):e0123245. Epub 2015/04/11. doi: 10.1371/journal.pone.0123245 PMID: 25860872; PubMed Central PMCID: PMC4393119.

22. Bai Y, Lan F, Yang W, Zhang F, Yang K, Li Z, et al. sRNA profiling in Aspergillus flavus reveals differentially expressed miRNA-like RNAs response to water activity and temperature. Fungal genetics and biology: FG & B. 2015; 81:113–9. Epub 2015/03/31. doi: 10.1016/j.fgb.2015.03.004 PMID: 25813270.

23. Dahlmann TA, Kuck U. Dicer-Dependent Biogenesis of Small RNAs and Evidence for MicroRNA-Like RNAs in the Penicillin Producing Fungus Penicillium chrysogenum. PloS one. 2015; 10(5):e0125989. Epub 2015/05/09. doi: 10.1371/journal.pone.0125989 PMID: 25955857; PubMed Central PMCID: PMC4425646.

24. Lau SK, Chow WN, Wong AY, Yeung JM, Bao J, Zhang N, et al. Identification of microRNA-like RNAs in mycelial and yeast phases of the thermal dimorphic fungus Penicillium marneffei. PLoS neglected tropical diseases. 2013; 7(8):e2398. Epub 2013/08/31. doi: 10.1371/journal.pntd.0002398 PMID: 23991243; PubMed Central PMCID: PMC3749987.

25. Chen Y-W, Lin C-J. Combining SVMs with various feature selection strategies. Feature extraction: Springer; 2006. p. 315–24.

26. Jiang P, Wu H, Wang W, Ma W, Sun X, Lu Z. MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. Nucleic acids research. 2007; 35 (Web Server issue):W339–44. Epub 2007/06/08. doi: 10.1093/nar/gkm368 PMID: 17553836; PubMed Central PMCID: PMC1933124.

27. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random forest: a classification and regression tool for compound classification and QSAR modeling. J Chem Inf Comput Sci. 2003; 43 (6):1947–58. Epub 2003/11/25. doi: 10.1021/ci034160g PMID: 14632445.

28. Yao Y, Ma L, Jia Q, Deng W, Liu Z, Zhang Y, et al. Systematic characterization of small RNAome during zebrafish early developmental stages. BMC Genomics. 2014; 15:117. Epub 2014/02/11. doi: 10.1186/1471-2164-15-117 PMID: 24507755; PubMed Central PMCID: PMC3932949.

29. Schneider TD, Stephens RM. Sequence logos: a new way to display consensus sequences. Nucleic acids research. 1990; 18(20):6097–100. Epub 1990/10/25. PMID: 2172928; PubMed Central PMCID: PMC332411.