CRF: detection of CRISPR arrays using random forest

2   Kai Wang[1] and Chun Liang[1]

3   [1] Department of Biology, Miami University, Oxford, OH 45056, USA

5   Corresponding Author:

6   Chun Liang[1]

7   501 E. High St, Oxford, OH 45056, USA

8   Email address: liangc@miamioh.edu

ABSTRACT

CRISPRs (clustered regularly interspaced short palindromic repeats) are particular repeat
sequences found in wide range of bacteria and archaea genomes. Several tools are available for
detecting CRISPR arrays in the genomes of both domains. Here we developed a new web-based
CRISPR detection tool named CRF (**C**RISPR Finder by **R**andom **F**orest). Different from other
CRISPR detection tools, a random forest classifier was used in CRF to filter out invalid CRISPR
arrays from all putative candidates and accordingly enhances detection accuracy. In CRF,
particularly, triplet elements that combine both sequence content and structure information were
extracted from CRISPR repeats for classifier training. The classifier achieved 94.42% accuracy
and 93.99% sensitivity. Moreover, CRF offers a highly interactive web interface for robust data
visualization that is not available among other CRISPR detection tools. After detection, the
query sequence, CRISPR array architecture, and the sequences and secondary structures of
CRISPR repeats and spacers can be visualized for visual examination and validation. CRF is
freely available at http://bioinfolab.miamioh.edu/crf/home.php.

INTRODUCTION

Found in bacteria and archaea genomes, CRISPRs (clustered regularly interspaced short
palindromic repeats) are special DNA sequences that are arrays of direct repeats (23-55 *nt*)
separated by different spacers with similar length. The first description of CRISPRs was in 1987
(Ishino et al. 1987). Acting as an acquired immune system, CRISPR arrays can embed foreign
DNA fragments into spacers and defend host prokaryotic genomes against the invasion of
viruses or plasmids (Barrangou et al. 2007). Accordingly, some of spacer sequences can
perfectly match to viruses or plasmids genomes. CRISPR arrays can be classified into three
systems based on the mechanism of inducing adaptive immunity (Hsu et al. 2014). In both Type
I and Type III CRISPR systems, the CRISPR RNAs (crRNAs) are generated from cleavage of
pre-crRNAs by CRISPR associated ribonucleases. These crRNAs can bind to certain proteins to
form RNA-protein complexes that can degenerate invading DNAs. In Type II CRISPR system, a
pre-crRNA pairs with a *trans*-activating CRISPR RNA (tracrRNA) to form a RNA-RNA duplex,
which can be identified and cleaved by certain nucleases. After cleavage, the resultant crRNA-
tracrRNA duplex can bind to Cas9 protein to form a protein complex that can degenerate
invading viruses or plasmid DNAs. As one example of Type II system, CRISPR/Cas9 currently
is a popular tool for genome editing. With artificial crRNAs, the CRISPR-Cas9 complex can be
used to accurately knockout target genes in eukaryotic genomes for genome editing. In order to
design efficient crRNA to knock out target genes, detection of CRISPR arrays is the first
essential step (Ran et al. 2013).
Currently, several tools are available for detecting CRISPR arrays. CRT (Bland et al. 2007)
was implemented in JAVA and PILER-CR (Edgar 2007) in C++ as executable programs.
Meanwhile, three other tools were implemented as web-based tools: CRISPRFinder (Grissa et al.
2007b), CRISPI (Rousseau et al. 2009), and CRISPRDetect (Biswas et al. 2016). All these
existing programs use either predefined parameters for seed region detection and extension in

49 repeat finding or a scoring system for CRISPR identification. For instance, PILER-CR finds
50 CRISPR candidates through local alignments of a query genome and uses sequence similarity
51 and length requirements as its criteria to filter out some invalid CRISPR arrays (Edgar 2007).
52 Differently, CRT deploys the strategy of finding *k-mer* seed regions and conducting sequence
53 extension in repeat detection. After obtaining all putative candidates, CRT applied simple criteria
54 of sequence similarity and length requirement to exclude invalid CRISPR arrays (Bland et al.
55 2007). In CRISPRFinder (Grissa et al. 2007b), Vmatch (Kurtz 2003) was used to find potential
56 CRISPR candidates first, then a scoring system was applied to score all candidates, and lastly
57 CRISPR candidates that meet certain score and length requirements were selected as valid ones
58 after eliminating tandem repeats using ClustalW (Thompson et al. 1994). As the most recent tool
59 for CRISPR detection, CRISPRDetect also utilizes a *k-mer* based seed region detection and
60 extension strategy, scores all candidates using a complex set of criteria, and reports those
61 candidates with high scores as the detected CRISPR arrays (Biswas et al. 2016). CRT, PILER-
62 CR and CRISPRFinder are powerful in detecting CRISPR arrays and are very easy to use.
63 However, their filtration steps are not accurate enough to exclude all invalid CRISPR arrays, as
64 demonstrated in our data analysis. Our study also revealed that many CRISPR arrays detected by
65 other tools are missed by CRISPRDetect, and many larger integral CRISPR arrays were split into
66 and reported as smaller CRISPR arrays by CRISPRDetect (see Supplemental Data).

67   Differentiating valid CRISPR arrays from invalid ones is very challenging. CRISPR
68 candidates with canonical repeat architecture might not be real or valid due to other sequence
69 features (e.g., secondary structure properties) that might affect DNA-RNA-protein interaction
70 (Jackson et al. 2014). As the first tool in CRISPR detection for archaea and bacterial genomes
71 that utilizes machine learning, we developed a new bioinformatics pipeline named CRF
72 (**C**RISPR Finder by **R**andom **F**orest). Innovatively, we applied a random forest classifier to
73 identify the real CRISPR arrays based on repeat sequence and structural features. Instead of
74 using a scoring system to differentiate real and invalid ones, the classifier was used to exclude
75 invalid CRISPR arrays from all candidates. 32 triplet elements that contain both sequence
76 content and structure information were extracted from both known and candidate CRISPR arrays
77 sequences for training and testing the classifier. The classifier achieved a very high accuracy
78 (94.42%) and sensitivity (93.99%), indicating that our classifier is powerful enough to
79 differentiate valid and invalid CRISPR arrays. In CRF, tandem repeat checking was also applied
80 to filter invalid CRISPR candidates. The whole pipeline was implemented in a web-based tool by
81 using PHP and PERL scripts. CRF is able to detect CRISPR arrays and provides a highly
82 interactive data visualization of the query sequence, array architecture, repeat weblogo, and the
83 sequences and secondary structures of both repeats and spacers, etc.

84 **MATERIALS AND METHODS**

85   CRF contains three parts as shown in Figure 1. First, C̲RISPR r̲ecognition t̲ool (CRT) was
86 used to detect all CRISPR array candidates. CRT was a widely used tool in finding CRISPR
87 arrays, and showed a better performance than PILER-CR (Edgar 2007). However, the default
88 parameters of CRT are very strict and fail to detect some valid CRISPR array candidates. In

89   order to detect more array candidates, we adjusted the parameters of CRT and particularly
90   changed the seed region of repeats from 8 *nt* to 5 *nt*. Since the repeat sequences of CRISPR
91   arrays could contain mismatches (Sorek et al. 2013) and the seed region is first used to search
92   perfect matches and then extended to tolerate mismatches in repeat detection, a shorter seed
93   region can enhance search sensitivity and allow more CRISPR candidates bearing mismatches
94   detected. For nucleotide extension at the 3' end of the identical seeds, the default threshold was
95   also changed from 0.75 to 0.6, while the 5'-end extension threshold was the default (i.e., 0.75) in
96   CRT. This means that we can extend the repeat seed region one nucleotide right (3') if at least
97   60%, rather than 75%, of repeats share the same nucleotide in this position. The reason why we
98   modified the threshold of 3' end extension, rather than 5' end extension, is that the 3' ends of
99   CRISPR repeats appear to have more variations than their 5' ends (Jiang et al. 2013). Obviously,
100 such a change in the extension threshold allows for detection of more candidate CRISPR arrays.
101   Second, we trained a random forest classifier to differentiate valid CRISPR repeats from
102 invalid repeat candidates. All CRISPR arrays from CRISPRdb were downloaded and the repeat
103 sequences from these known CRISPR arrays were extracted for training. After removing the
104 redundancy, 11407 repeats were used as the positive dataset. The negative dataset was the
105 random sequences generated from the positive dataset by using a first-order Markov model.
106 There were 12000 random sequences within the negative dataset. Sequence content and
107 secondary structure analyses of direct repeats in CRISPR arrays reveal that some CRISPRs have
108 stable stem-loop structures due to the palindromic nature of their repeats (Kunin et al. 2007).
109 Thus, we extracted the triplet elements of repeat sequences as feature vectors to train, classify
110 and predict repeat sequences. Such triplet elements prove to be very discriminative in identifying
111 short sequences possessing stem-loop structures because these kinds of features contains both
112 sequence content and structure information (Wang et al. 2014; Xue et al. 2005). As shown in
113 Figure 2, 32 triplet elements can be used to represent each repeat sequence effectively. In CRF,
114 the secondary structures of repeats were predicted by RNAfold (Hofacker et al. 1994). The
115 random forest model was trained using randomForest (version 4.6-10) R package (Liaw &
116 Wiener 2002). The positive and negative datasets were randomly divided into a training set (80%)
117 and a testing set (20%) as shown in Table 1. The random forest classification model was trained
118 with the default parameters (i.e. ntree is 500, mtry is 5, and the threshold of splitting criterion is
119 50%). The input sequences will be labeled either "repeat" (i.e., CRISPR repeat) or "random"
120 after the classification.
121   Repeats, spacers, and whole CRISPR arrays need to be checked to exclude potential tandem
122 repeats that are unlikely to be true CRISPR arrays (Skennerton et al. 2013). We used Phobos
123 (Mayer 2008) to identify potential tandem repeats among all CRISPR candidates we detected.
124 Phobos gives each candidate a score for being a tandem repeat, and we used the default threshold
125 score of 10 for tandem repeat determination. Moreover, a *k-mer* (k=3) check was applied to
126 exclude potential tandem repeats. The default *k-mer* (k=3) threshold values for repeats, spacers
127 and whole CRISPR arrays are 0.2, 0.2, and 0.25 respectively. These threshold values were
128 defined based on the known CRISPR arrays from CRISPRdb. Users can modify these values
129 through "Advance Settings" on the home page of CRF to get more specific results. Furthermore,

130 in order to exclude invalid CRISPR arrays, Hamming distance was calculated to see if a CRISPR
131 array contains same-length spacers. Among all spacers in the CRISPR array, if any two spacers
132 have a similarity larger than 50%, then the CRISPR array candidate will be discarded (Barrangou
133 & Marraffini 2014).

RESULTS

Random Forest Classifier

We used a random forest classifier to differentiate real CRISPR repeats from invalid ones using triplet elements as sequences features. After the training and testing processes, the model achieved accuracy (94.42%) and sensitivity (93.99%). This high accuracy indicates that our classifier is able to differentiate real CRISPR repeats from invalid CRISPR repeats. The **A**rea **U**nder **C**urve (AUC) of this classifier is 0.9837. This high AUC value suggests that this classifier has great performance at CRISPR repeat identification. Each CRISPR array candidate will be classified as a valid or invalid CRISPR array based on their repeat sequences.

CRF web service

The web interface of CRF was implemented in PHP and JavaScript whereas PERL was used for its backend pipeline. The web service consists of three pages ("*Home*", "*Manual*", and "*Contact*"). The query sequence should be one in FASTA format and ambiguous nucleotide N characters are accepted. Users can use the default settings or click the "*Advanced Settings*" button to modify parameters for CRF. These parameters are mainly about changing the sequence length of repeats and spacers and adjusting values to remove tandem repeats from the CRISPR candidates detected.

After submitting a query sequence on the CRF website, the final valid CRISPR arrays are shown in a summary table. On the top of the results page, the query sequence can be displayed and visualized in SeqViewer by clicking the "*Show sequence in SeqViewer*" button. In SeqViewer, users can search sequence motifs by using the motif search input field on the navigation bar. The use of IUPAC codes is acceptable for the motif search function, which enables powerful fuzzy search of motifs. For the detailed results of each detected CRISPR array, a user can click the ID/Name on the result list to jump to the relevant position on the result page. The detailed results of each CRISPR array consists of four parts: (1) the basic information of a CRISPR array is listed at the beginning of the detailed results, (2) a SVG-based CRISPR array architecture graph and the sequence follows as the second part, where a user can click the sequence to show secondary structure of each repeat or spacer, (3) the sequence logo figure of CRISPR repeats obtained through WebLogo 3.4 (Crooks et al. 2004) is displayed next, and (4) the web link that allows a user to download relevant figures and data is provided last. Users can download the spacers or the whole CRISPR array as a sequence file in FASTA format.

Comparison with other tools

For performance comparison, we compared CRF with PILER-CR, CRISPRDetect and CRT, we only choose these three programs for comparison because CRISPRFinder and CRISPI are web-based tool that do not offer source codes for local machine installation. We tested these four

169 different programs using the same genome dataset that has been utilized for random forest
170 classifier testing. The non-redundant repeat sequences reported by CRISPRdb that were used as
171 the testing dataset in our study were extracted from 1139 genomes. These 1139 genomes were
172 used for comparison. All four programs were executed with their default parameters and settings.
173 The results were compared to CRISPRdb, because CRISPRdb is the most comprehensive, widely
174 utilized up to date CRISPR database (Grissa et al. 2007a). Based on the detected CRISPR arrays,
175 we found that PLIER-CR can detect more arrays that begin or end at wrong positions in
176 comparison with the known CRISPRs from CRISPRdb (Bland et al. 2007). CRISPRDetect
177 showed better results, but it also missed certain portions of CRISPR arrays. CRT is better than
178 other the two tools, but it cannot filter out tandem repeat sequences.

179   The result shows that 541 out of 1139 genomes can be detected with CRISPR arrays by all
180 of these four programs (Fig. 3). Individually, PILER-CR, CRT, CRISPRDetect and CRF
181 detected CRISPR arrays in 1033, 1067, 604 and 948 out of 1139 genomes, respectively. In
182 CRISPRdb, those arrays with only one spacer flanked by two direct repeats are also maintained
183 as "questionable structures" (Bland et al. 2007), whereas a minimum of two spacers with three
184 direct repeats are required for being valid CRISPR arrays by all four programs. This explains
185 why all four programs detected few genomes with CRISPR arrays. Comparing these four
186 programs, PILER-CR, CRT and CRF detected CRISPR arrays in 80% of 1139 genomes whereas
187 CRISPRDetect detection rate was only 50%. Clearly, this indicates that CRISPRDetect missed a
188 significant portion of CRISPR arrays, in comparison with the other three programs.

189   In CRISPRdb, 3689 CRISPR arrays are distributed in 1139 testing genomes. We used these
190 CRISPR arrays as references to calculate sensitivity and precision for these four programs. The
191 results are shown in Table 2. CRF, CRT, and PILER-CR achieved above 60% sensitivity (*i.e.*
192 CRF: 61.6%, CRT: 66.2%, and PILER-CR: 69.7%). CRISPRDetect only achieved 36.5%
193 sensitivity, and this indicates that CRISPRDetect missed almost half of the CRISPR arrays in
194 these testing genomes. For precision, CRF achieved 83.3% while the other programs only
195 achieved 73.8% (CRT), 51.9% (CRISPRDetect), and 75.2% (PILER-CR). This result suggests
196 that our program is more effective in excluding fake CRISPR arrays than other programs.

197   CRISPR repeats are defined as short palindromic repeats, and the length and palindromic
198 structure of the repeats are crucial for interaction between CRISPR RNAs and Cas proteins
199 (Nishimasu et al. 2014). However, the repeats do not show an obvious palindromic structure in
200 Type II and Type III CRISPR systems (Hsu et al. 2014). Therefore, we focused on the length
201 distribution of the repeats from detected CRISPR arrays to compare the four programs. The
202 CRISPR repeats from the CRISPR arrays detected from the test data set by all four programs
203 were extracted for length distribution comparison. As shown in figure 4, it shows that CRT,
204 CRISPRDetect and CRF have a similar distribution pattern located in repeat length ranged from
205 23 *nt* to 40 *nt*, which is consistent with that of CRISPRdb. In CRISPRdb, because a strict cutoff
206 (minimum repeat length equals 23 *nt*) was applied, no repeat shows a length in the range from 20
207 to 23 *nt*. In our CRF pipeline, we retained the CRISPR arrays that have repeat lengths located in
208 the range of 20 to 23 *nt*, since CRISPR arrays with repeat lengths of 20 to 23 *nt* have been
209 reported by R. Jansen (Jansen et al. 2002). The CRISPR repeats from PILER-CR show a

210 different distribution pattern in which the repeats are longer than the repeats detected by the
211 other three programs. Although PILER-CR can detected more CRISPR arrays than other three
212 programs, Figure 4 supports the finding that CRISPR arrays from PILER-CR are more likely to
213 start or end at wrong positions in comparison with CRT and PATSCAN (Bland et al. 2007). In
214 terms of length distribution, CRT, CRISPRDetect and CRF show better results PILER-CR.
215 However, CRT cannot filter out the tandem repeat sequences in its reported CRISPR arrays (See
216 supplement files). In our CRF pipeline, Phobos was used to remove tandem repeats candidates.
217 For longer repeats ranging from 44 to 50 *nt*, CRISPRDetect detected nothing in this range while
218 CRF, CRISPRdb, and CRT showed a small peak value of around 47 *nt*. Considering the fact that
219 only 604 genomes were detected with CRISPR arrays by CRISPRDetect, it is clear that that
220 CRISPRDetect still fails to detect a noticeable portion of CRISPR arrays. Our data analysis
221 suggests that CRF is a better program for CRISPR array detection due to its consistency with
222 CRISPRdb regarding repeat length distribution applying machine learning to reduce invalid
223 candidates and using an effective strategy to filter out tandem repeats.

224 **CONCLUSIONS**

225 Currently, CRISPR-Cas system is one of the most popular genome editing tools. In order to
226 design effective CRISPR RNA, detecting CRISPR arrays from bacterial and archaea genomes is
227 an important step. Several tools are available for detecting CRISPR arrays. As a valuable
228 addition, we designed a new pipeline called CRF to detect CRISPR arrays from genome data, in
229 which we utilized machine learning and tandem repeat filters. Using 1139 genomes, we
230 compared CRF with other three tools CRT, PILER-CR, and CRISPRDetect. Clearly,
231 CRISPRDetect failed to detect a significant portion of CRISPR arrays, PILER-CR found more
232 CRISPR arrays at wrong end or start positions, and CRT cannot filter out tandem repeat
233 sequences. Our comparison indicates that CRF shows better performance than the other
234 programs. Moreover, differently from the other tools, CRF offers a highly interactive web
235 interface useful for data visualization, sequence examination and CRISPR architecture validation.
236

237 **ACKNOWLEDGEMENTS**

238 We thank Matthew Spinelli and Joshua Hoeksema for comments and suggestions to improve this
239 manuscript.
240

241 **REFERENCES**

242 Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, and Horvath P. 2007. CRISPR
243   provides acquired resistance against viruses in prokaryotes. *Science* 315:1709-1712.
244 Barrangou R, and Marraffini LA. 2014. CRISPR-Cas systems: Prokaryotes upgrade to adaptive immunity. *Mol Cell*
245   54:234-244.
246 Biswas A, Staals RH, Morales SE, Fineran PC, and Brown CM. 2016. CRISPRDetect: A flexible algorithm to define
247   CRISPR arrays. *BMC Genomics* 17:356.

248 Bland C, Ramsey TL, Sabree F, Lowe M, Brown K, Kyrpides NC, and Hugenholtz P. 2007. CRISPR recognition tool
249   (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC*
250   *Bioinformatics* 8:209.
251 Crooks GE, Hon G, Chandonia JM, and Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Res*
252   14:1188-1190.
253 Edgar RC. 2007. PILER-CR: fast and accurate identification of CRISPR repeats. *BMC Bioinformatics* 8:18.
254 Grissa I, Vergnaud G, and Pourcel C. 2007a. The CRISPRdb database and tools to display CRISPRs and to generate
255   dictionaries of spacers and repeats. *BMC Bioinformatics* 8:172.
256 Grissa I, Vergnaud G, and Pourcel C. 2007b. CRISPRFinder: a web tool to identify clustered regularly interspaced
257   short palindromic repeats. *Nucleic Acids Res* 35:W52-57.
258 Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, and Schuster P. 1994. Fast folding and comparison of
259   RNA secondary structures. *Monatshefte für Chemie/Chemical Monthly* 125:167-188.
260 Hsu PD, Lander ES, and Zhang F. 2014. Development and applications of CRISPR-Cas9 for genome engineering. *Cell*
261   157:1262-1278.
262 Ishino Y, Shinagawa H, Makino K, Amemura M, and Nakata A. 1987. Nucleotide sequence of the iap gene,
263   responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene
264   product. *J Bacteriol* 169:5429-5433.
265 Jackson RN, Golden SM, van Erp PB, Carter J, Westra ER, Brouns SJ, van der Oost J, Terwilliger TC, Read RJ, and
266   Wiedenheft B. 2014. Structural biology. Crystal structure of the CRISPR RNA-guided surveillance complex
267   from Escherichia coli. *Science* 345:1473-1479.
268 Jansen R, Embden J, Gaastra W, and Schouls L. 2002. Identification of genes that are associated with DNA repeats
269   in prokaryotes. *Mol Microbiol* 43:1565-1575.
270 Jiang W, Maniv I, Arain F, Wang Y, Levin BR, and Marraffini LA. 2013. Dealing with the evolutionary downside of
271   CRISPR immunity: bacteria and beneficial plasmids. *PLoS Genet* 9:e1003844.
272 Kunin V, Sorek R, and Hugenholtz P. 2007. Evolutionary conservation of sequence and secondary structures in
273   CRISPR repeats. *Genome Biol* 8:R61.
274 Kurtz S. 2003. The Vmatch large scale sequence analysis software. *Ref Type: Computer Program*:4-12.
275 Liaw A, and Wiener M. 2002. Classification and regression by randomForest. *R news* 2:18-22.
276 Mayer C. 2008. Phobos, a tandem repeat search tool for complete genomes. *Version*.
277 Nishimasu H, Ran FA, Hsu PD, Konermann S, Shehata SI, Dohmae N, Ishitani R, Zhang F, and Nureki O. 2014. Crystal
278   structure of Cas9 in complex with guide RNA and target DNA. *Cell* 156:935-949.
279 Ran FA, Hsu PD, Wright J, Agarwala V, Scott DA, and Zhang F. 2013. Genome engineering using the CRISPR-Cas9
280   system. *Nature protocols* 8:2281-2308.
281 Rousseau C, Gonnet M, Le Romancer M, and Nicolas J. 2009. CRISPI: a CRISPR interactive database. *Bioinformatics*
282   25:3317-3318.
283 Skennerton CT, Imelfort M, and Tyson GW. 2013. Crass: identification and reconstruction of CRISPR from
284   unassembled metagenomic data. *Nucleic Acids Res* 41:e105.
285 Sorek R, Lawrence CM, and Wiedenheft B. 2013. CRISPR-mediated adaptive immune systems in bacteria and
286   archaea. *Annual review of biochemistry* 82:237-266.
287 Thompson JD, Higgins DG, and Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple
288   sequence alignment through sequence weighting, position-specific gap penalties and weight matrix

289   choice. *Nucleic Acids Res* 22:4673-4680.

290   Wang K, Liang C, Liu J, Xiao H, Huang S, Xu J, and Li F. 2014. Prediction of piRNAs using transposon interaction and

291   a support vector machine. *BMC Bioinformatics* 15:1-8.

292   Xue C, Li F, He T, Liu G-P, Li Y, and Zhang X. 2005. Classification of real and pseudo microRNA precursors using local

293   structure-sequence features and support vector machine. *BMC Bioinformatics* 6.

294

Figure 1

The workflow of C RISPR Finder by Random Forest (CRF) Pipeline.

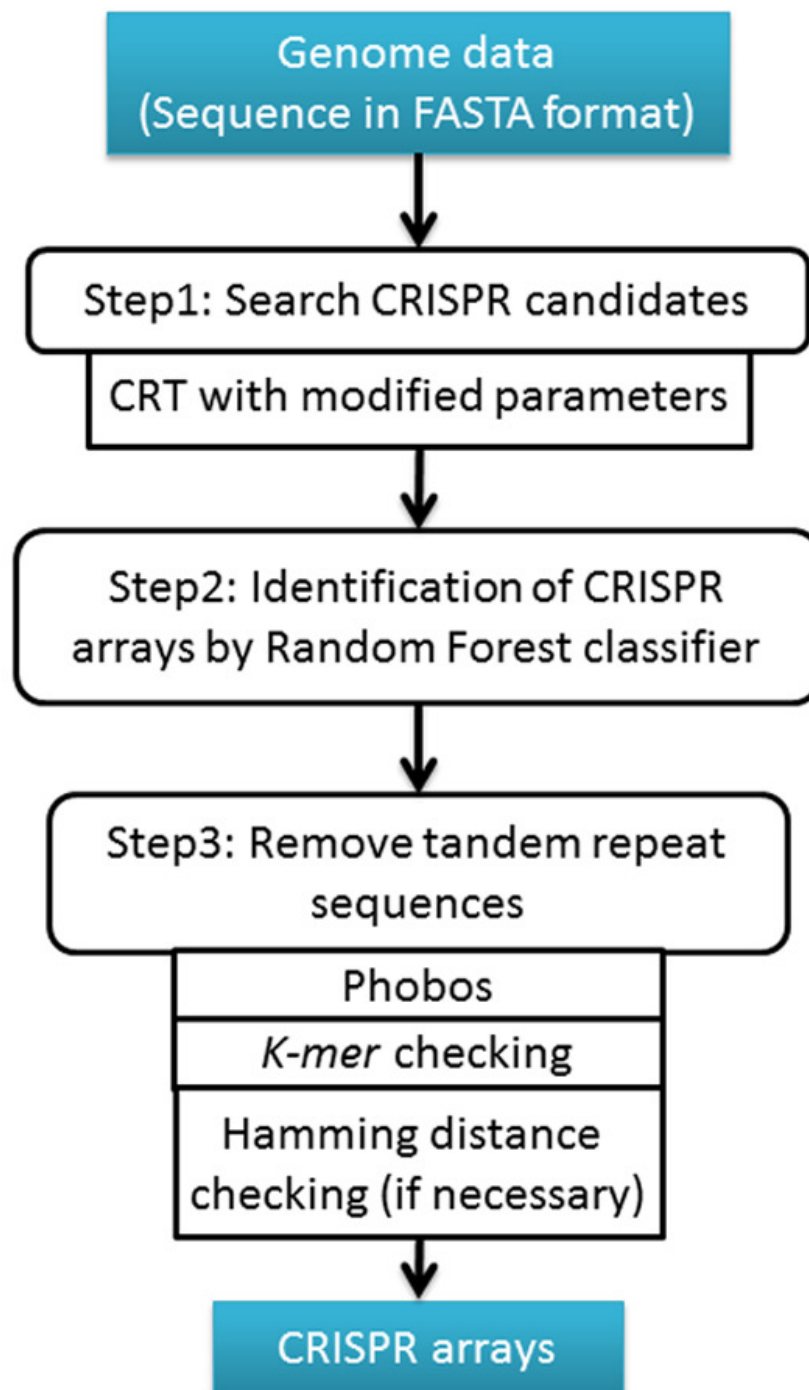



Figure 2

Principle of triplet elements that contain both sequence and structural information.

The secondary structure was predicted by RNAfold as shown in (a). The triplet elements consist of the dot-bracket structure and the sequence content. (b) A moving window (length = 3) will slide over from the left end to the right end of the dot-bracket structure, and a sub-structure with the length of 3 will be extracted each time. Each sub-structure combined with the middle nucleotide of the three adjacent nucleotides will be counted as one triplet element. (c) Because there are only two structural statuses for each nucleotide (paired/bracket versus unpaired/dot), in total, three adjacent nucleotides will have 8 different structural combinations (e.g., "(((", "((.", "(..", "(.(", ".((", ".(.", "..(", and "..."). With "A", "G", "C", and "T" nucleotides, there are 32 triplet elements that have both sequence content (i.e., nucleotide) and structural information. So for each sequence, the number of 32 triplet elements will be counted and normalized through dividing by the total element number. The normalized data was shown in (d).

(a)

(b)

```
...((((((..........))))))...
CAGACCGTAAGTTAGCGCTACGGTTCA
```

A	G	A		C
...	..(	.((		...

(c)

32 triplet elements

```
(((A,  ((.A,  (.(A,  .((A,
(..A,  .(.A,  ..(A,  ...A,
(((G,  ((.G …
```

(d)

32 vector features

```
0.1999, 0.3333, 0.4555,
0, 0.1363, 0.0454, 0, …
```

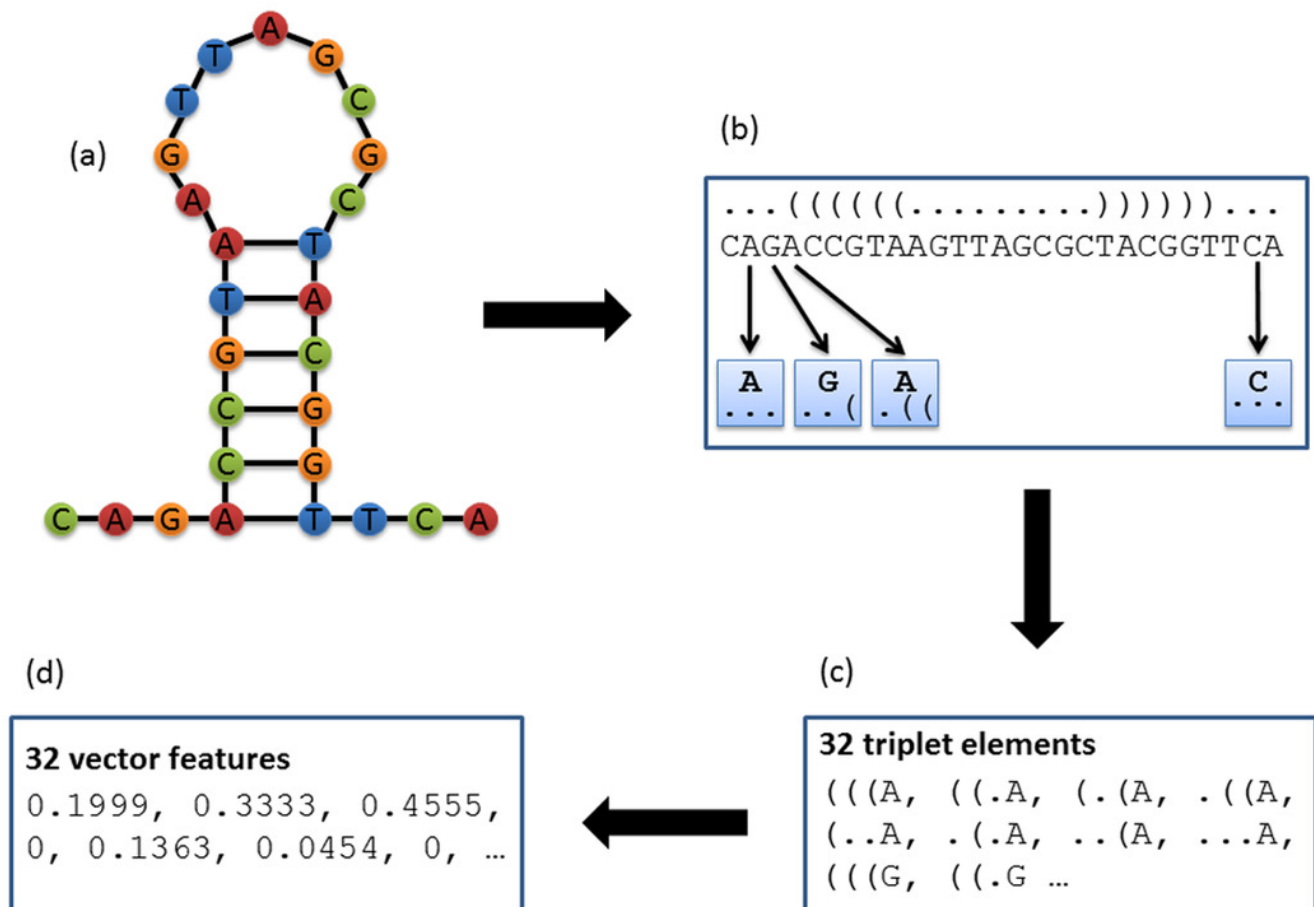



Figure 3

Comparison of the number of genomes detected with CRISPR arrays by four programs.

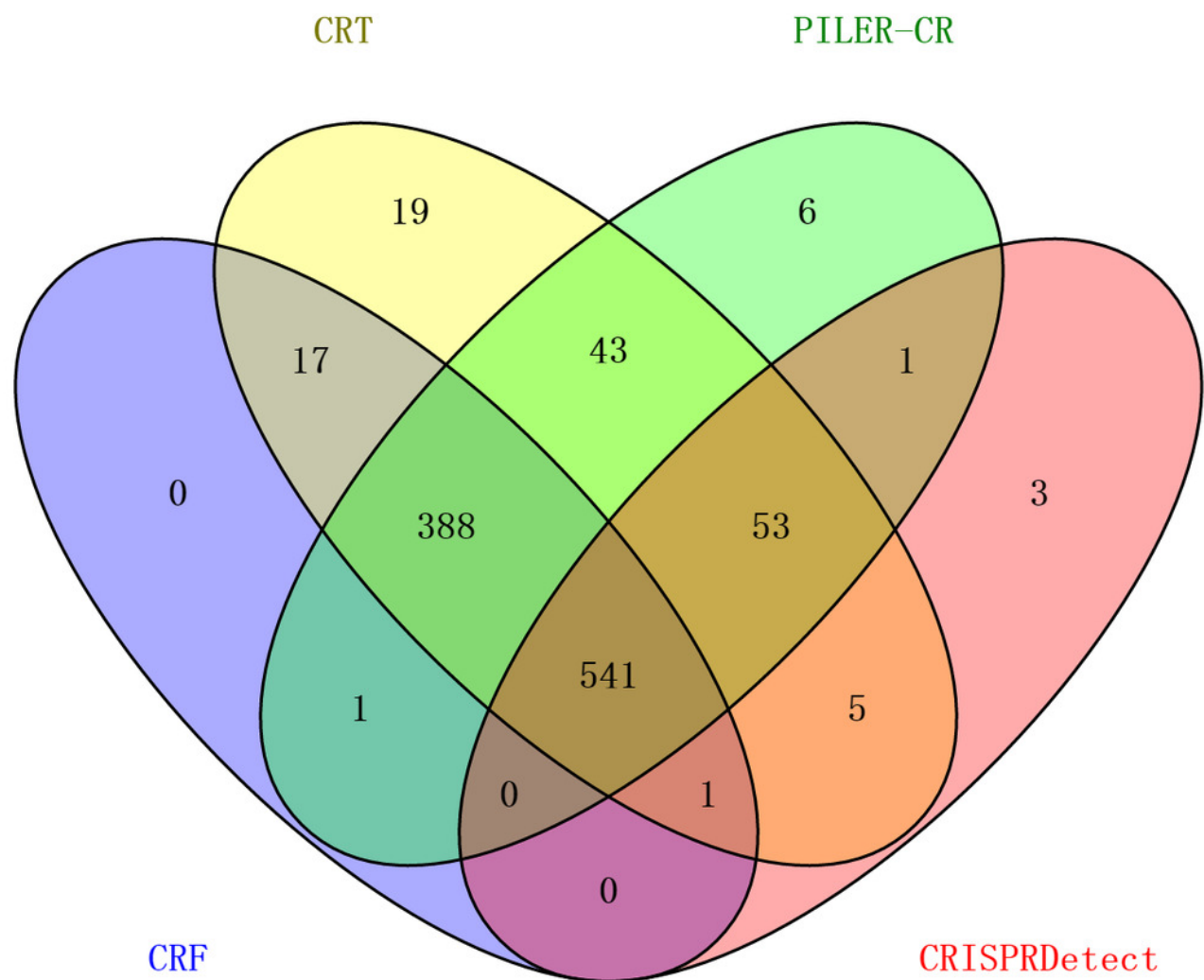
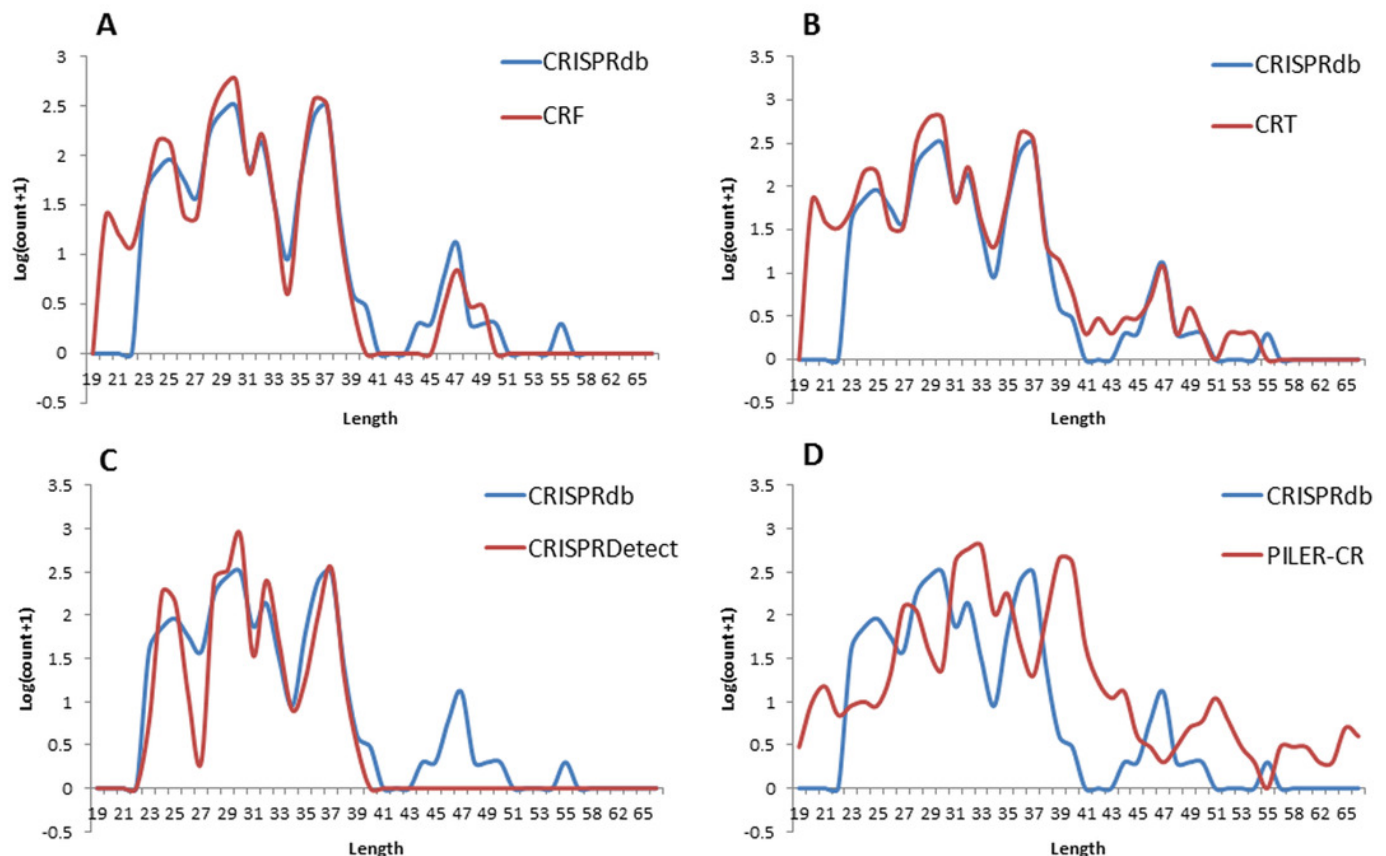



Figure 4

Repeat length distribution of CRISPR repeats detected by four programs, in comparison with the complete CRISPR repeats annotated in CRISPRdb.

Datasets for random forest classifier training and testing.

	CRISPR repeats	Random sequences
Training set	9,126	9,600
Testing set	2,281	2,400
Total number	11,407	12,000

Comparison of sensitivity and precision between four programs.

	CRF	CRT	CRISPRDetect	PILER-CR
Sensitivity (%)	61.62	66.28	34.46	69.75
Precision (%)	**83.29**	73.80	51.93	75.21