Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

Genome analysis

CRISPR/Cas9 cleavage efficiency regression through boosting algorithms and Markov sequence profiling

Hui Peng [1], Yi Zheng [1], Michael Blumenstein [1], Dacheng Tao [2] and Jinyan Li [1,*]

[1] Advanced Analytics Institute, Faculty of Engineering and IT, University of Technology Sydney, PO Box 123, Broadway, NSW 2007, Australia [2] School of Information Technologies and the Faculty of Engineering and Information Technologies, University of Sydney, J12/318 Cleveland St, Darlington, NSW 2008, Australia

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: CRISPR/Cas9 system is a widely used genome editing tool. A prediction problem of great interests for this system is: how to select optimal single guide RNAs (sgRNAs) such that its cleavage efficiency is high meanwhile the off-target effect is low.

Results: This work proposed a two-step averaging method (TSAM) for the regression of cleavage efficiencies of a set of sgRNAs by averaging the predicted efficiency scores of a boosting algorithm and those by a support vector machine (SVM). We also proposed to use profiled Markov properties as novel features to capture the global characteristics of sgRNAs. These new features are combined with the outstanding features ranked by the boosting algorithm for the training of the SVM regressor. TSAM improved the mean Spearman correlation coefficiencies comparing with the state-of-the-art performance on benchmark datasets containing thousands of human, mouse and zebrafish sgRNAs. Our method can be also converted to make binary distinctions between efficient and inefficient sgRNAs with superior performance to the existing methods. The analysis reveals that highly efficient sgRNAs have lower melting temperature at the middle of the spacer, cut at 5'-end closer parts of the genome and contain more 'A' but less 'G' comparing with inefficient ones. Comprehensive further analysis also demonstrates that our tool can predict an sgRNA's cutting efficiency with consistently good performance no matter it is expressed from an U6 promoter in cells or from a T7 promoter in vitro.

Availability: Online tool is available at http://www.aai-bioinfo.com/CRISPR/. Python and Matlab source codes are freely available at https://github.com/penn-hui/TSAM.

Contact: Jinyan.Li@uts.edu.au

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

CRISPR/Cas9 (the clustered, regularly interspaced, short palindromic repeats/CRISPR-associated protein 9 system) is a widely used genome editing tool. The system can be reprogrammed by changing the sequence of its single-guide RNA (sgRNA) for site-specific cutting of the target DNA strand (Mali *et al.*, 2013; Shalem *et al.*, 2014; Bolukbasi *et al.*, 2016), applicable for the investigation of gene functions (Swiech *et al.*, 2015), gene expressions (Konermann *et al.*, 2015), and clinical trials (Yin *et al.*, 2017). CRISPR/Cas9 is mainly composed of a Cas9 protein and an sgRNA as a complex. It is the 20nt-long spacer sequence in the sgRNA that can induce the site-specific binding of the CRISPR/Cas9 complex to its target genome locus located at the upstream of a protospacer adjacent motif (PAM 'NGG', where 'N' can be 'A', 'G', 'C' or 'T'). The key to good design of sgRNAs is to determine the spacer sequence by selecting a protospacer sequence complementary with the spacer's target sequence such that the cleavage (cleaving) efficiency is high.

There are two critical prediction problems in the selection of sgRNAs. The first problem is the prediction of whether the sgRNA on-target cleaving efficiency is high or not. The subsequent problem is whether the sgRNA's off-target effect is low (Fu *et al.*, 2013; Shen *et al.*, 2014; Kleinstiver *et al.*, 2016). The first question is fundamental. Our work here focuses on machine learning algorithms for assessing the cleaving efficiencies of candidate sgRNAs. The algorithms make regressions on the numerical values of their cleavage efficiencies. The algorithms can be also turned to make binary classifications between high-efficiency and low-efficiency sgRNAs. The second question about the sgRNA off-target effects is closely linked to the first one. As it involves genome-wide number of genes and some experimental methods such as GUIDE-seq (Tsai *et al.*, 2015) and Digenome-seq (Kim *et al.*, 2015), the off-target prediction problem will be investigated separately.

Prediction algorithms have been recently proposed to identify efficient sgRNAs through characterizing their spacer sequence preferences (Doench *et al.*, 2014; Xu *et al.*, 2015; Wong *et al.*, 2015; Kaur *et al.*, 2016; Moreno-Mateos *et al.*, 2015), thermodynamics features (Doench *et al.*, 2014; Wong *et al.*, 2015) and structure features (Wong *et al.*, 2015). The sequence features are widely adopted because many nucleotide preference phenomena have been observed. For example, nucleotides distal to the PAM were found to be dominated by the guanine enrichment, while the remaining nucleotides are characterized by the cytosine enrichment (Moreno-Mateos *et al.*, 2015). These nucleotide preference properties have been exploited to differentiate efficient sgRNAs from those inefficient ones by machine learning methods such as support vector machine (SVM) (Doench *et al.*, 2014; Wong *et al.*, 2015; Kaur *et al.*, 2016; Rahman and Rahman, 2017). In particular, a regression method (Doench *et al.*, 2016) has been proposed to predict the numerical values of the cleaving efficiencies for candidate sgRNAs. Its novel idea is a Rule Set 2 (RS2) for predicting the on-target activities of sgRNAs. Different from the previous classification methods, this regression model also uses some new features such as cutting position features and the two nucleotides in the N and N positions relative to the PAM 'NGGN'. Though RS2 achieved remarkable performance, there still exists large space for improving the performance.

We introduce a two-step averaging method (TSAM) for the prediction of sgRNA cleaving efficiencies. At the first step, a boosting regression model is trained on the conventional feature space of sgRNAs to map these sgRNAs to their cleaving efficiency scores. At the second step, we use Markov sequence profiles of sgRNAs as new features together with important features selected by the boosting algorithm to train a non-linear SVM to make regression again on the cleaving efficiencies. The two scores are then averaged as the predicted cleaving efficiencies of these sgRNAs.

Both the boosting algorithm and the Markov sequence profiling have the same aim to exploit important characteristic features of sgRNAs to improve the prediction performance but at different aspects. Literature methods already proposed a large number of features to describe sgRNAs. However, not all of them are effective for the prediction of the cleavage efficiencies. The newly introduced Markov sequence features can capture the global sequence characteristics of sgRNAs which are different from the conventional position-specific preferences (Doench *et al.*, 2014; Wong *et al.*, 2015; Kaur *et al.*, 2016; Doench *et al.*, 2016). The boosting algorithm, XGBoost (Chen and Guestrin, 2016), is a scalable end-to-end tree boosting system that can rank the feature importance during the training process. XGBoost is also a state-of-the-art regression algorithm with better performance than the traditional gradient boosting trees (Doench *et al.*, 2016), having a wider range of applications (Zhang *et al.*, 2017; Torlay *et al.*, 2017). Furthermore, our two-step averaging strategy underlines a complementary nature of the boosting regression approach and the SVM regression approach. From our experiments, the regression results of XGBoost and SVM are always different. It is good to integrate the two regression results to improve the prediction performance on the sgRNA cleaving efficiencies.

Markov sequence profiles of an sgRNA are extracted through a profile Hidden Markov Model (pHMM). It works by converting a multiply sequence alignment for sequences from a known family into a position-specific scoring system (Eddy, 1998). This system can be used to evaluate whether a new sequence is a homologous sequence of this sequence family. This method has been leveraged to address many other biological sequence related bioinformatics problems (Karplus *et al.*, 1998; Schliep *et al.*, 2003; Wheeler *et al.*, 2013; Huo *et al.*, 2017). In this work, sgRNA sequences are first grouped into sub-families in accordance with their efficiency scores. Then, probabilities of a given sgRNA being a homologous sequence for each sub-families are formed as a multi-pHMM vector for characterizing the global features of sgRNA sequences. An SVM regressor trained with only pHMM properties can obtain similar mean Spearman correlations comparing with the state-of-the-art methods. Hence, we decided to combine pHMM features with the top-ranked features of XGBoost to train the second-step SVM regressor for a better performance.

The performances of our TSAM is compared with the state-of-the-art regression methods such as RS2 (Doench *et al.*, 2016) and CRISPRscan (Moreno-Mateos *et al.*, 2015). On Doench's FC dataset (human and mouse sgRNAs), TSAM obtained a mean Spearman correlation of 0.583, better than RS2's 0.522. On the RES dataset (human sgRNAs) and the FC+RES dataset, TSAM achieved mean Spearman correlations 0.530 and 0.567 respectively, better than RS2's 0.455 and 0.510. On the dataset which was used by CRISPRscan containing 1020 zebrafish sgRNA sequences, TSAM can achieve a competitive Pearson correlation of 0.49 (comparing with CRISPRscan's 0.45). Our two-step regression approach was converted into a binary classification method to distinguish between high-efficiency and low-efficiency sgRNAs. The classification performance on the benchmark datasets also outperforms the state-of-the-art methods. For instance, the mean AUC of the three-fold cross validation on Xu's ribosomal dataset (Xu *et al.*, 2015) is 0.896, much exceeding Xu's 0.843. For the cross-gene validation and cross-platform validation, our performance are 0.813 and 0.840 respectively, better than Xu's 0.778 and 0.757.

Haeussler *et al.* (2016) advised that the performance of an on-target efficiency prediction model is strongly dependable on whether the guide RNA is expressed from an U6 promoter or it is transcribed in vitro with the T7 promoter. To compare the performance of TSAM with the state-of-the-art methods on datasets of different expression systems, we collected abundant datasets from Haeussler *et al.* (2016) to test two specified versions of TSAM. One is named TSAM_U6, which was trained with the FC+RES dataset as input, in which the guide RNAs are all transcribed from the U6 promoter. The dataset CRISPRScan containing guides expressed from the T7 promoter in vitro was used to build the second predictor named TSAM_T7. The results confirmed that our TSAM can always achieve better performances on both of the U6 and T7 promoter datasets.

Our case studies are related to the optimal sgRNAs selected for gene therapies to cure the retinitis pigmentosa and X-linked chronic granulomatous disease (Yu *et al.*, 2017; De Ravin *et al.*, 2017). The highly efficient sgRNAs recommended by our method can well match with those sgRNAs which had been validated by wet lab experiments and domain experts. This partly proves the effectiveness of our prediction tool, and illustrates the great potential of our method for practical use.

2 Materials and methods

2.1 High throughput genome engineering datasets for building the regression and classification models

We tested the algorithms on total 11 datasets. Three datasets from (Doench *et al.*, 2016) were downloaded to build our TSAM regression model. The three datasets are named: the FC dataset which contains 1841 sgRNAs with the flow cytometry (FC) method detecting the knockdowns; the RES dataset which contains 2549 sgRNAs with their knockdown efficiencies measured through drug resistance detection; and the combined dataset (FC+RES). We removed 10 sgRNAs from the FC dataset because of their ambiguous mapping to the reference genome (Fusi *et al.*, 2015). Doench's paper reported that there are 1831 curated sgRNAs in the FC dataset, however, there are only 1830 unique sgRNAs from their supplementary materials. Furthermore, 1020 sgRNAs for cleaving zebrafish genome sequences were acquired from (Moreno-Mateos *et al.*, 2015). Different from FC and RES, where the guides are transcribed from U6 promoters in cells, this zebrafish dataset contains the guides expressed from T7 promoters in vitro. As the cutting efficiency measurement methods are distinct, separate models are trained and evaluated on these different datasets. More details of these four datasets are listed at the first 4 rows of **Table 1**.

Table 1. 11 datasets for construction and evaluation of our classification and regression models

Name	validation type	sample size	literature
FC	logo[a]	1830	(Doench *et al.*, 2016)
RES	logo	2549	(Doench *et al.*, 2016)
FC+RES	logo	4379	(Doench *et al.*, 2016)
CRISPRScan	ShuffleSplit	1020	(Moreno-Mateos *et al.*, 2015)
Xu_ribo	threefold	731H,438L[b]	(Xu *et al.*, 2015)
Xu_non-ribo	inter-geneset[c]	671H,237L	(Xu *et al.*, 2015)
Xu_mouse	inter-platform[d]	830H,234L	(Xu *et al.*, 2015)
Xu_inde1	independent[e]	52H,25L	(Xu *et al.*, 2015)
Xu_inde2	independent	110H,110L	(Xu *et al.*, 2015)
Chari_spCas9	tenfold	133H,146L	(Chari *et al.*, 2015)
Chari_stlCas9	tenfold	82H,69L	(Chari *et al.*, 2015)

[a] regression, leave-one-gene-out cross-validation

[b] classification, where H for efficient and L for inefficient

[c] trained on Xu_ribo and tested on Xu_non-ribo

[d] trained on Xu_ribo + Xu_non-ribo and tested on Xu_mouse

[e] trained on Xu_ribo + Xu_non-ribo + Xu_mouse and tested on Xu_inde1

In the test of whether our TSAM can address the problem of classifying sgRNAs into high-efficiency or low-efficiency ones, five datasets from (Xu *et al.*, 2015) were downloaded including three datasets for three-fold cross validation, inter-geneset validation and inter-platform validation, and two independent test sets (directly from the authors) for evaluation and comparing the performances of different methods. The details are listed at the 5th to 9th rows of **Table 1**.

To compare with Chari's sgRNA Scorer (Chari *et al.*, 2015), their datasets were obtained from the supplementary files of the published paper (shown at the last two rows of **Table 1**). Chari et al. tested their method on two datasets: a 133 high-activity vs 146 low-activity sgRNA dataset for the assessment of spCas9 system, and a 82 high vs 69 low sgRNA dataset for the stlCas9 system (from *Streptococcus thermophilus*, where its PAM is NNAGAAW).

2.2 Features for building the regression and classification models

2.2.1 Conventional sequence features

Here, an sgRNA sequence is always referred to as the protospacer sequence corresponding to the spacer and its upstream to the PAM. To extract some similar features as used by RS2 (Doench *et al.*, 2016), we similarly extended the sequences to 30nt in length, namely $N_4N_{20}NGGN_3$ (N represents any nucleotide, the first 4nt and the last 3nt are also extracted together with the original 20nt spacer and the PAM NGG). An sgRNA sequence is denoted as $S = s_1s_2s...s_i...s_{30}$, where $s_i \in \{A, G, C, T\}$.

Nucleotide composition features: The number of each single nucleotide (e.g., how many 'A') in S is counted, and each characterized as an order 1 nucleotide composition (nc1) feature. Similarly, the number of each dinucleotides or trinucleotides (e.g., how many 'AA' or 'AAA' in S) is counted, and each characterized as an order 2 or order 3 nucleotide composition feature (nc2, or nc3). The counts of the dinucleotides and the trinucleotides were computed by a sliding window mechanism.

Position specific nucleotide binary features: An order 1 position specific nucleotide binary feature (psnb1), at a given position, is initialized as a vector (0, 0, 0, 0). The first element represents whether the nucleotide at this position is 'A'. If yes, change the 0 to be 1. The second element represents the status of 'G', the third for 'C' and the forth for 'T'. For example, if at position 1, the nucleotide is 'A' then, this vector is (1, 0, 0, 0), or if the nucleotide is 'C', this vector is (0, 0, 1, 0). Similarly, an order 2 position specific nucleotide binary feature (psnb2) and order 3 position specific nucleotide binary feature (psnb3) are established in the same way, where every dinucleotide and trinucleotide are used as an element of the 16-dimensional vector and 64-dimensional vector at a given position.

GC features: Each of these features describes the counts of how many 'G' or 'C' in S (named GC counts features), or the percentage of 'G'+'C' in S (named the GC percent feature).

2.2.2 Thermodynamic features

The melting temperatures of sgRNA sequences at different regions were computed with the Biopython Tm_staluc function (Cock *et al.*, 2009; Le Novere, 2001). We considered the following regions as features: the whole 20nt spacer (TMr1), the core region (12nt adjacent to PAM, TMr2), the non-core region (the remaining 8nt of the 20nt spacer, TMr3), the whole 30nt extended sgRNA sequence (TMr4), the 5nt adjacent to PAM (TMr5), the 8nt proximal to the previous 5nt (TMr6) and another 5nt next to the middle 8nt (TMr7). The last four regions have been used by RS2 (Doench *et al.*, 2016).

2.2.3 Cutting position related features

Cutting positions relative to protein sequences have been used to improve the performance on the prediction of sgRNA cleaving efficiencies (Doench *et al.*, 2016). In this work, we considered the cutting position to the genome sequence (cut_geno), to the transcript sequence (cut_trans) and to the protein sequence (cut_pro) as three features. Meanwhile, the percentage of the cutting length was considered as a feature computed as the length from the start of the sequence to the cut position divided by the whole sequence length (denoted as cut_per_geno, cut_per_trans, and cut_per_pro respectively). The gene's genome sequence, transcript sequence, protein sequence and the detail exon, intron, 5'UTR and 3' UTR sequences were downloaded from the ensembl database (Hubbard *et al.*, 2002) for the mapping of these cutting positions. The gene's start coordination was normalized to be 1 for calculating feature values of cut_geno, cut_trans, cut_per_geno and cut_per_trans. Features cut_trans, cut_pro, cut_per_pro and cut_per_trans were set to be a value of 0 if the sgRNA cut in an intron region. Features cut_pro and cut_per_pro were also set to be a value of 0 if the sgRNA cut at non-coding regions.

2.2.4 Profile hidden Markov model (pHMM) features of sgRNA sequences

It is the sgRNA sequence as a whole that can truly determine its cutting efficiency. Here, the global features of an sgRNA sequence are extracted through a profile hidden Markov model (Eddy, 1998). We hypothesized that those sgRNAs with similar cutting efficiencies should contain more sequence similarities, and vice versa. Thus, these sgRNAs can be grouped into subfamilies where the efficiencies of the sgRNAs in each group are similar. Then, if a new sequence belongs to a subfamily, its cutting efficiency may also similar to its homologous sequences. The pHMM was adopted to solve this homologous sequence searching problem, where the pHMM properties were used to characterize the sgRNA sequences.

A pHMM is usually used for modeling multiple sequence alignments and it can provide a probabilistic model for comparing new sequences to the multiple alignments (Durbin *et al.*, 1998). Traditional pHMM can be described with an HMM composed of a state set $S = \{Begin, Match, Insert, Delete, End\}$ and an alphabet of symbols $\mho = \{e_1, e_2, \dots\}$ that are emitted by the non-silent states (usually are Match and Insert states). After training on a sequence family (a protein family or a set of homologous gene sequences), a transition probability matrix and an emission probability matrix can be constructed to depict the transitions between the states and the emission status of the non-silent states. For a given sequence, a log-sum-of-odds score describing the probability of the pHMM generating it can be computed by the *Viterbi* algorithm (Forney, 1973). Please be referred to (Eddy, 1998; Durbin *et al.*, 1998) for more details about pHMM.

Most of the high throughput experiments fixed the spacer length as 20nt. Thus, the spacer sequences here were set to be well aligned with the fixed length 20 (there is no Insert or Delete state but only Match status), where the pHMM is a so-called BLOCK-style ungapped motif (Eddy, 1998). Two sets of symbols were permitted to be emitted at the Match state, i.e., a single nucleotide set $\mho_1 = \{A, G, C, T\}$ and a dinucleotide set $\mho_2 = \{AA, AG, AC, AT, \dots, TA, TG, TC, TT\}$. To avoid the emission probability of zero, we add pseudocounts into the observed counts. Therefore, the emission probability e_i is calculated as $e_M(e_i) = \frac{count(e_i)+pu}{count(all)+pd}$, where, pu and pd are the pseudocounts for the observed count of each emitted symbol and all the emissions.

Suppose there is a set of sgRNAs $Sg = \{sg_1, sg_2, \dots, sg_j, \dots, sg_m\}$ with known efficiencies $Ef = \{ef_1, ef_2, \dots, ef_j, \dots, ef_m\}$, $ef_j \in [0, 1]$. For an sgRNA ℓ, its pHMM properties are extracted by the following two steps:

- **Step1: Grouping Sg into k sub-families and training their pHMMs.** Separating Sg into k sub-families $Sf = \{sf_1, sf_2, \dots, sf_x, \dots, sf_k\}$, where each of them has an efficiency range, e.g., $ef(sf_x) \in [0.1, 0.2)$. For $sf_x \in Sf$ and a given emission symbol type t, a pHMM can be trained with its sequences. These pHMMs are denoted as $H^t = \{h_1^t, h_2^t, \dots, h_x^t, \dots, h_k^t\}$.
- **Step2: Extracting ℓ's pHMM vector.** For sgRNA ℓ, the probability hf_x^t generated by h_x^t is computed by the *Viterbi* algorithm, and ℓ is characterized by a vector $Hf_\ell^t =< hf_1^t, hf_2^t, \dots, hf_x^t, \dots, hf_k^t >$, where $t = \mho_1, \mho_1$.

Here both of the two emission symbol sets $\mho_1$ and $\mho_2$ are used, which can produce two vectors for sgRNA ℓ, i.e., $Hf_\ell^{\mho_1}$ (pHMMe1) and $Hf_\ell^{\mho_2}$ (pHMMe2).

2.3 Procedures for training our TSAM

Our TSAM cleaving efficiency regression model is built by four main steps. Firstly, all the features are created. Then, an XGBoost regressor is trained with some selected primary features to estimate the first-step scores. The features' importance are evaluated simultaneously. Later, the most important features are combined with the pHMM features to optimize a RBF SVM regressor. Then the second-step scores are calculated. At last, the first-step scores and the second-step scores are averaged as the final scores for the regression. **Figure 1** shows the flowchart to construct TSAM.

To get the best training performance on the dataset FC, the XGBoost and SVM regression methods were both optimized by the leave-one-gene-out cross-validation for the best parameters. The best parameters were fixed when these two regression methods were used to generate leave-one-gene-out cross-validation performance on the RES dataset or on the FC+RES dataset. To have a fair performance comparison with Moreno-Mateos *et al.* (2015) on the CRISPRScan dataset, our regression methods were also optimized by the same Shuffle-Split cross-validation as Moreno-Mateos *et al.* (2015) did.

We also note that there is a pre-evaluation process to select important features from the initial feature set for optimizing the XGBoost regressor. This process is implemented by the backward elimination strategy (Mao, 2004) with default parameters for XGBoost. During each fold of the cross-validation, the selected features are assigned with feature importance to weight their contributions for optimizing the regressor.

The features that work well for SVM (e.g., the pHMMe1 and pHMMe2 according to our results) are combined with the boosting selected top-K important features to train a RBF kernel SVM regressor (libSVM v3.22 (Chang and Lin, 2011)). As the features' importance are evaluated during each cross-validation fold, the final selected important features are the union of the top-K ones from all the folds. This SVM regressor predicts the second-step scores for the sgRNAs. The details of determining the parameters for regressors and features are described in the **Supplementary file 1**.

3 Results

We first report the cleavage preferences of sgRNAs as revealed by XGBoost and explain how these preferences are different from literature observations. Then, we report excellent regression performance achieved by integrating XGBoost and SVM. These results and analysis are mainly focused on the dataset FC. After that, we present comparison results between our method and the state-of-the-art methods to demonstrate the superior performance on the sgRNA cleavage efficiency regression by our method. At last, two case studies are presented to illustrate the effectiveness of our method for practical use in gene therapies.

3.1 Nucleotide and cleavage preferences of highly efficient sgRNAs as revealed by the boosting algorithm

Some interesting nucleotide preferences of the highly efficient sgRNAs are revealed by the XGBoost algorithm on the FC dataset (see **Figure 2**). A highly efficient sgRNA is always a sequence of relatively lower melting temperature at the middle of the spacer, in comparison with those of low efficiencies (a mean value 8.84 for the highly efficient sgRNAs that are ranked at the top 20% of the 1830 sgRNAs according to their actual efficiencies vs 13.11 for the low efficient sgRNAs ranked at the bottom 20%, p-value=1.04E-09 under the two-sample Kolmogorov-Smirnov test (Lilliefors, 1967)). Also, the highly efficient sgRNAs prefer to cut at the 5'-end closer part of a gene (a mean value of cut_per_geno is 41.56% for the highly efficient sgRNAs vs 46.61% for the low efficient sgRNAs, p-value=1.51E-04). In addition, the nucleotide composition of the highly efficient sgRNAs and the low efficient sgRNAs exhibits a distinct divergence: the highly efficient sgRNAs have more 'A' (on average 6 for the highly efficient sgRNAs vs. 5 for the low efficient ones, p-value=2.83E-09), but less 'G' (on average 10 for the highly efficient vs. 11 for the low efficient, p-value=5.06E-03), 'GG' (on average 4 for the highly efficient vs. 5 for the low efficient, p-value=3.23E-08) and 'GGG' (on average 1 for the highly efficient vs. 2 for the low efficient, p-value=6.51E-07).

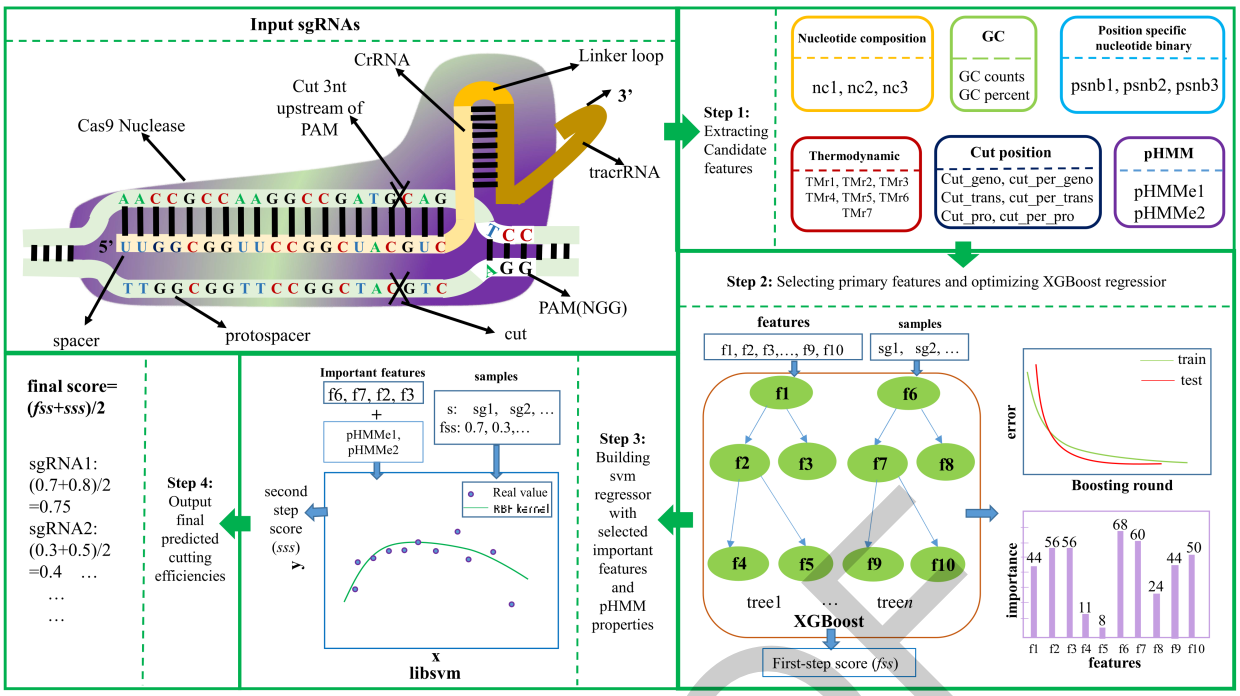

Fig. 1. The flowchart to construct TSAM for predicting sgRNA cleavage efficiencies. This flowchart contains four main steps: at first 6 types of initial features are created; in the second step, primary features are selected from the initial feature set to optimize an XGBoost regressor and output the first-step scores (fss) and the importance scores of the features; then, the important features are combined with the pHMM features to train a RBF kernel SVM and compute the second-step score (sss); lastly, the first-step score and the second-step score of an sgRNA is averaged as the final predicted score ($(fss + sss)/2$).

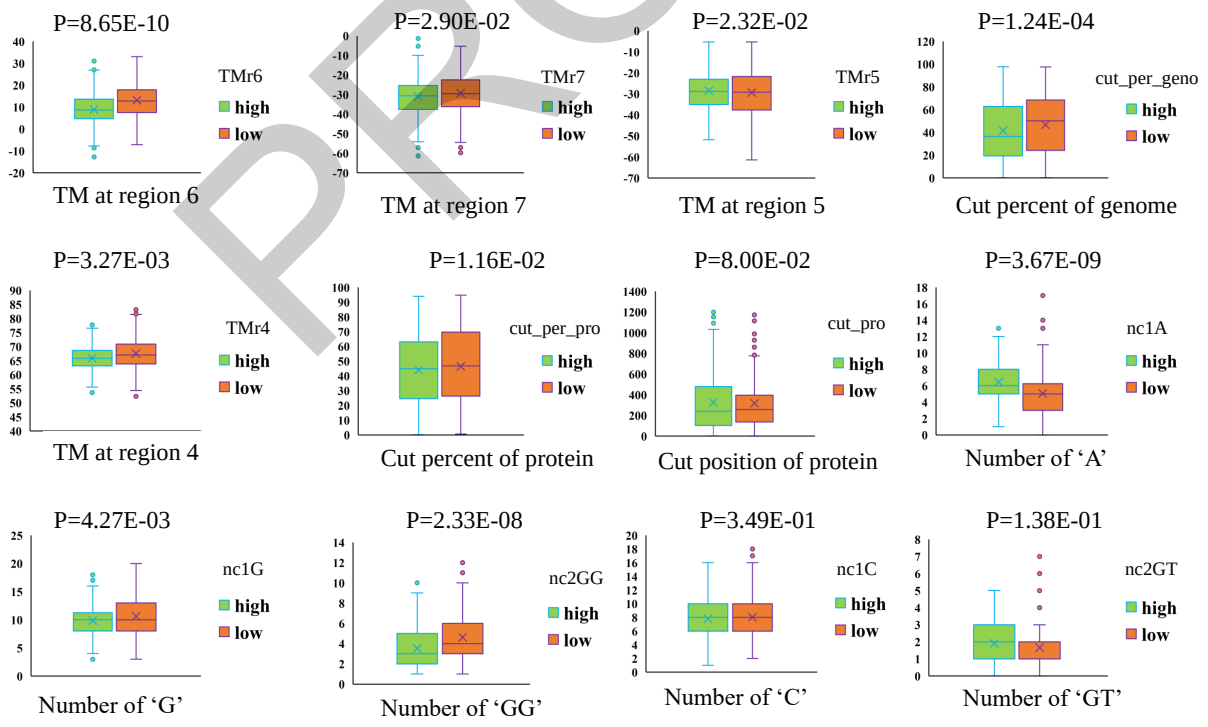

Fig. 2. Top 12 important features and analysis on the nucleotide and cleavage preferences. Y-axis shows the feature values. The feature names are placed under the x-axis and their symbols are placed at the top right panel of the subplots. These features are ranked by their importance. Type "high" means that the sgRNAs are ranked at top-20% while the "low" represents that the sgRNAs are ranked at bottom-20%. The p-value shown in each sub-figure is computed via the two-sample Kolmogorov-Smirnov test.

Doench *et al.* (2016) have reported that the three types of features that contribute substantially to the efficiency prediction are: position-independent counts of single and dinucleotides, location of the sgRNA within the protein, and melting temperatures at different regions (having Gini importance of 16%, 13% and 11% respectively). By our boosting algorithm, these three types of features constitute the top 25 sub-features whose importance are higher than 100. Different conclusions are drawn as follows. First, the melting temperatures at different regions are the best features (with a mean importance 542.64), then the cutting position related features are (with a mean importance 424.41), followed by the nucleotide composition related features (with an average importance 136.59). Meanwhile, the cutting percent relative to genome DNA sequence (cut_per_geno, not applied by RS2) is better than the cutting percent relative to protein (cut_per_pro) and the cutting position at the protein (cut_pro) (importance are 503.89, 399.44 and 369.89 respectively). The divergences of the values for cut_per_pro and cut_pro between the high and low efficient sgRNAs are not as significant as that of the cut_per_geno (p-value=1.39E-02, 8.41E-02 and 1.51E-04 respectively).

The regression performance on the cleaving efficiencies by our XGBoost is better than Doench et al.'s RS2. We obtained a mean Spearman correlation 0.562, but RS2 obtained only 0.522 on the FC dataset. This is why conclusions on the nucleotide preferences of highly efficient sgRNAs are different between these two methods. We note that our XGBoost regressor did not use all the features but only important features such as TMr4-TMr7, nc1, nc2, nc3, psnb1, psnb2, GC counts, GC percent, cut_per_geno, cut_pro and cut_per_pro (form 677 dimensions in total). More details about the XGBoost regression parameter settings and the features can be found at our **Supplementary file 1**.

3.2 Further performance improvement by integrating pHMM properties

The pHMM properties (combining the pHMMe1 and pHMMe2) can be used to build an SVM regressor to achieve fairly good performance, where a mean Spearman correlation 0.519 was obtained. Adding the top ranked important features evaluated by the former boosting can further improve the SVM regressor's mean Spearman correlation to 0.559 which is superior to Doench's methods (RS2's mean Spearman correlation=0.522 and L1-Regression's mean Spearman correlation=0.513). If the pHMM properties were removed from this strong SVM regressor, the performance dropped about 0.01. This implies that the pHMM properties are indispensable to construct our excellent SVM regressor.

The proposed TSAM obtained a mean Spearman correlation 0.583 which is much better than Doench's methods. It also improves the mean Spearman correlation of our XGBoost regressor by 0.021, benefited from its integration with the SVM regressor trained on the pHMM properties and other significant features. The SVM regressor alone also achieved better performance than Doench's methods but worked not as well as TSAM. This proves that the XGBoost regressor and the SVM regressor can predict the sgRNA's cutting efficiencies cooperatively. The parameter optimization process is described in **Supplementary file 1**.

3.3 Results on 11 benchmark datasets comparing with the state-of-the-art methods

Four benchmark datasets were used to evaluate the performance of our proposed TSAM. The performance was compared with the following stat-of-the-art methods: Doench et al's RS2, L1-Regression methods (implemented by this work) (Doench *et al.*, 2016), and the CRISPRscan method (Moreno-Mateos *et al.*, 2015). Our TSAM improves the mean Spearman correlation by more than 0.05 comparing with RS2 and L1-Regression on the FC, RES and the FC+RES datasets (under the leave-one-gene-out evaluation framework), and improves the mean Pearson correlation by about 0.04 comparing with CRISPRscan (under the same Shuffle-Split evaluation framework) on the sgRNAs dataset for cutting zebrafish genome sequences. The detailed results are presented at the first four rows of **Table 2**.

Table 2. Regression performance of different methods on four benchmark datasets

Methods	regression performance on			
	FC	RES	FC+RES	CRISPRscan
RS2	0.522	0.455	0.510	-
L1-Regression	0.513	0.468	0.505	-
CRISPRscan	-	-	-	0.45
TSAM	**0.583**	**0.530**	**0.567**	**0.488**
TSAM-MT1	0.565	0.441	0.531	0.475
TSAM-MT2	0.575	0.493	0.555	0.477

In the further evaluation of TSAM, we have conducted cross-dataset test. We trained TSAM on the FC dataset, and then the sgRNAs belonging to the 8 genes in the RES dataset were adopted as 8 independent test sets. The mean Spearman correlation by our regression is 0.431, which is much better than the performance by Doench's methods (0.397 by RS2 and 0.383 by the L1 regression). On the 8 genes, we obtained higher Spearman correlations on 6 of them than Doench's RS2 and L1 regression methods. When TSAM was trained with the 2549 sgRNAs from the RES dataset, and tested on the 9 genes from the FC dataset, the mean Spearman correlation was 0.551 for TSAM, while Doench's RS2 and L1 regression obtained only 0.508 and 0.493 respectively. As expected, we obtained better Spearman correlations than Doench's methods on 7 of the 9 genes.

We have conducted a stricter performance evaluation for TSAM to satisfy practical use conditions especially assuming the cutting position features are not accessible. For this performance test, we modified TSAM as two Mutation Types (MT): TSAM-MT1 and TSAM-MT2. TSAM-MT1 was trained without cutting position features (674-d, deleting the cut_per_geno, cut_pro and cut_per_pro), and TSAM-MT2 was trained without the cutting position related to the protein features (675-d, without cut_pro and cut_per_pro). The performances of these two variant methods are shown in the last two rows of **Table 2**. It is understood that the cutting position features can significantly affect the performance of our TSAM on the RES dataset. Except for one case testing on the RES dataset, our methods obtained much better performance than the state-of-the-art methods.

Our TSAM regression method can be easily converted for a binary classification approach to the distinction between highly efficient sgRNAs and low efficient ones. The steps are as follows. First, XGBoost was optimized to output feature importance scores (classification with the binary logistic function). Then, the important features were combined with the pHMM properties to train an SVM classifier with a RBF kernel (the pHMM group is set as 2, such as positive sample group and the negative sample group, probabilities as output). Then the classifier was tested on 7 datasets including 5 datasets for cross-validation and 2 independent test sets. The other classifiers (Doench *et al.*, 2014; Xu *et al.*, 2015; Chari *et al.*, 2015) were also optimized with the corresponding validation types in **Table 1**. The cross-validations were repeated 10 times and the performances were averaged as the final performance. Then the classification performances were weighted by Matthews correlation coefficient (MCC) (Matthews, 1975), F1, AUC and Accuracy which are all shown in **Table 3**.

Table 3. Performance comparison between our method and the state-of-the-art methods for the binary classification of sgRNAs

Method	dataset	MCC	F1	AUC	Accuracy
TSAM	Xu_ribo	0.640	0.871	0.896	0.834
Xu et al.'s	Xu_ribo	-	-	0.843	-
TSAM	Xu_non-ribo	0.505	0.884	0.813	0.822
Xu et al.'s	Xu_non-ribo	-	-	0.778	-
TSAM	Xu_mouse	0.508	0.891	0.840	0.830
Xu et al.'s	Xu_mouse	-	-	0.757	-
TSAM	Xu_inde1	0.311	0.800	0.798	0.714
Xu et al.'s	Xu_inde1	-	-	0.729	-
Doench et al.	Xu_inde1	-	-	0.648	-
TSAM	Xu_inde2	0.433	0.748	0.779	0.700
Xu et al.'s	Xu_inde2	-	-	0.711	-
Doench et al.'s	Xu_inde2	-	-	0.583	-
TSAM	Chari_spCas9	0.551	0.758	0.859	0.772
Chari et al.'s	Chari_spCas9	-	-	-	0.732
TSAM-MT1	Chari_stlCas9	0.718	0.865	0.930	0.855
Chari et al.'s	Chari_stlCas9	-	-	-	0.815

The variant method TSAM-MT1, instead of TSAM itself, was applied to test the performance on the Chari_stlCas9 dataset. The reason is that the PAM of the sgRNAs was defined as 'NNAGAAW' but not the 'NGG' motif. Thus the cutting position features could not be defined. We can see that TSAM-MT1 can outperform the state-of-the-art methods as well for the binary classification of sgRNAs. More comparison results are provided at **Supplementary file 1**.

3.4 Performance of TSAM on more datasets related to the U6 and T7 expression system

We used the datasets from Haeussler *et al.* (2016) to confirm that the proposed TSAM can work better than RS2 when the guide RNAs are expressed from U6 and better than CRISPRscan when the expression system is T7.

3.4.1 Comparison on datasets from the U6 expression system

We compared the prediction performance of TSAM_U6 and RS2 on 7 big datasets containing sgRNAs for cutting human or mouse genomes. Both TSAM_U6 and RS2 are trained on the FC+RES dataset, where the sgRNAs are expressed from U6 promoters in cells. The Spearman correlation are shown in **Table 4**.

We can see that for all the seven datasets each containing more than 1000 sgRNAs, our TSAM_U6 achieved about 3% more the Spearman correlation than RS2.

3.4.2 Comparison on datasets from T7 expression system

Another 5 datasets whose sgRNAs are expressed from T7 promoters were used to compare the performances between TSAM_T7 and CRISPRscan. Both of these two predictors were trained with the CRISPRScan dataset and the sgRNAs in this dataset are expressed from a T7 promoter in vitro. The Spearman correlations are listed in **Table 4**. Again, the proposed TSAM_T7 achieved 10% more the Spearman correlation on 3 out of 5 datasets and about 5% more on the remaining two datasets than the best existing predictor CRISPRscan for this type of expression system. See our **Supplementary file 1** and **Supplementary file 3** for detailed results and the applied datasets.

3.5 Case study: designing sgRNAs for gene therapy

Crispr/Cas9 system is a very promising genome engineering tool for curing genetic diseases (Men *et al.*, 2017). In the understanding of whether TSAM can recommend reasonable sgRNAs for practical use, we conducted case studies for recommending sgRNAs to treat retinitis pigmentosa and X-linked chronic granulomatous disease. Gene editing investigations on these two diseases have been successfully undertaken by domain experts recently (Yu *et al.*, 2017; De Ravin *et al.*, 2017).

Yu *et al.* (2017) attempted to knockdown gene Nrl to prevent retinal degeneration in a mouse model and suggested adopting CRISPR/Cas9-mediated NRL disruption in rods as a promising treatment option for patients with retinitis pigmentosa. For our prediction, the genome sequences of mouse Nrl gene was downloaded from Ensembl database under the transcript id ENSMUST00000062232.13. Total 138 potential spacer sequences were found with the PAM 'NGG'. Among these 138 candidate sgRNAs, the cleavage efficiencies of those sgRNAs cutting at the coding region were predicted by our TSAM method. If considering just the cutting efficiency, the 3 top-ranked sgRNAs' spacer sequences are 5'-ATGCCTGGCTCACTGAAGGT-3' (s1, cut efficiency=0.850), 5'-GTATGGTGTGGAGCCCAACG-3' (s2, cut efficiency=0.801) and 5'-CACAGACATCGAGACCAGCG-3' (s3, cut efficiency=0.762). Yu's work proposed to use 5 candidate sgRNAs (denoted NT1 to NT5). They finally selected NT2 as an optimal sgRNA because it contains relative higher ability to generate indels and lower predicted off-target potential. Our s2 exactly matches with their NT2 (in comparison, RS2 ranks this optimal sgRNA at the sub-optimal 3rd position, while CRISPRscan ranks it at the 28th position among all the potential sgRNAs for cutting Nrl). This suggests that our TSAM cleavage efficiency regression method is quite accurate for recommending good sgRNAs for disease gene editing. Our method is indeed useful to suggest only several top-ranked sgRNAs (e.g., top 3) for narrowing down the search scope in the subsequent filtering such as the off-target prediction and in vivo experimental test. Such a recommendation approach can save time and costs, meanwhile achieving satisfactory accuracy.

De Ravin *et al.* (2017) investigated a gene repair problem with Crispr/Cas9 to cure patients with X-linked chronic granulomatous disease that arises from mutations in CYBB (C676T substitution in exon 7 of CYBB gene). Different from the above case study, to correct the point mutation, the cutting site should be close to the mutation site. Four potential sgRNAs (gRNA1, gRNA2, gRNA3 and gRNA8) whose cutting sites are near the mutation site were tested. They found that gRNA2 (5'-CACCCAGATGAATTGTACGT-3') had the maximal cutting efficiency. By our TSAM (exactly, TSAM-MT1 is used, because these sgRNAs cut at non-coding regions), the predicted scores of the four sgRNAs are: 0.310 for gRNA1, 0.693 for gRNA2, 0.534 for gRNA3 and 0.243 for gRNA8. For comparison, the predicted scores by RS2 are quite differently as 0.364, 0.704, 0.555 and 0.351 respectively. On the other hand, CRISPRscan could detect just gRNA3 (score=28) and gRNA8(score=35), but not gRNA1 or gRNA2 (gRNA1 and gRNA2 start with 'TT' and 'CA' respectively, thus they cannot be expressed from the T7 promoter and predicted by CRISPRscan (Moreno-Mateos *et al.*, 2015)). Thus, TSAM can accurately recommend the optimal sgRNA for the mutation correction case as well.

4 Conclusion

We proposed a two-step averaging method (named TSAM) to conduct regressions on the cleavage efficiencies of sgRNAs. The first-step cleavage efficiency scores are predicted by an optimized XGBoost regressor. This step also ranks the features' importance for feature selection. At the second step, an SVM regression model is constructed using the pHMM features combined with the top-ranked features selected by the first step.

Table 4. Spearman correlation of TSAM, RS2 and CRISPRscan tested on datasets from U6 or T7 expression systems

U6 expression system					
dataset	size	genome	literature	TSAM_U6	RS2
Wang/Xu HL60	2076	*Mouse*	Wang *et al.* (2014)	**0.517**[a]	0.485
Chari 293T	1234	*Human*	Chari *et al.* (2015)	**0.382**	0.381
Hart Rpe	4214	*Mouse*	Hart *et al.* (2015)	**0.309**	0.281
Hart Hct116-2 Lib 1	4239	*Mouse*	Hart *et al.* (2015)	**0.416**	0.384
HartHelalib1	4256	*Mouse*	Hart *et al.* (2015)	**0.388**	0.353
HartHelalib2	3845	*Mouse*	Hart *et al.* (2015)	**0.394**	0.359
XuKBM	2076	*Mouse*	Xu *et al.* (2015)	**0.540**	0.512
T7 expression system					
dataset	size	genome	literature	TSAM_T7	CRISPRscan
Eschstruth Zebrafish	17	*Zebrafish*	Haeussler *et al.* (2016)	**0.224**	-0.0043
Varshney Zebrafish	102	*Zebrafish*	Varshney *et al.* (2015)	**0.363**	0.262
Gagnon Zebrafish	111	*Zebrafish*	Gagnon *et al.* (2014)	**0.410**	0.357
Shkumatava Zebrafish	162	*Zebrafish*	Haeussler *et al.* (2016)	**0.292**	0.258
Teboul Mouse In Vivo	30	*Mouse*	Haeussler *et al.* (2016)	**0.565**	0.426

[a] For each dataset, the highest Spearman correlation is in bold

The first score and the second score are averaged as the cleavage efficiency of each sgRNA in the prediction. Our regression method can be easily converted into a binary classification method for the distinction between high-efficiency sgRNAs and low-efficiency sgRNAs. TSAM was evaluated on 11 benchmark datasets containing thousands of sgRNAs editing human, mouse and zebrafish genome sequences and on additional 12 datasets of different expression system. The performance of TSAM was compared with the state-of-the-art methods to prove its superior performance. Two case studies have also demonstrated the effectiveness of TSAM. Our future work will focus on the integration of off-target prediction methods with the current on-target efficiency prediction algorithm to build a more comprehensive tool for sgRNA design where higher efficiency and specificity can be achieved simultaneously. In addition, more definitions of 'PAM' will be considered for TSAM. The cross-species cross-expression system performance evaluation will be investigated in near future when the supporting datasets are publicly available.

Acknowledgments

We thank Professors Xiaole Shirley Liu and John Doench for providing us their un-published datasets.

Funding

This work was supported in part by Australia Research Council research projects FT130101457 and DP140102164.

References

Bolukbasi, M. F., Gupta, A., and Wolfe, S. A. (2016). Creating and evaluating accurate CRISPR-Cas9 scalpels for genomic surgery. *Nature Methods*, **13**(1), 41–50.

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, **2**(3), 27.

Chari, R., Mali, P., Moosburner, M., and Church, G. M. (2015). Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach. *Nature Methods*, **12**(9), 823–826.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM.

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., *et al.* (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**(11), 1422–1423.

De Ravin, S. S., Li, L., Wu, X., Choi, U., Allen, C., Koontz, S., Lee, J., Theobald-Whiting, N., Chu, J., Garofalo, M., *et al.* (2017). CRISPR-Cas9 gene repair of hematopoietic stem cells from patients with X-linked chronic granulomatous disease. *Science Translational Medicine*, **9**(372), eaah3480.

Doench, J. G., Hartenian, E., Graham, D. B., Tothova, Z., Hegde, M., Smith, I., Sullender, M., Ebert, B. L., Xavier, R. J., and Root, D. E. (2014). Rational design of highly active sgRNAs for CRISPR-Cas9–mediated gene inactivation. *Nature Biotechnology*, **32**(12), 1262–1267.

Doench, J. G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E. W., Donovan, K. F., Smith, I., Tothova, Z., Wilen, C., Orchard, R., *et al.* (2016). Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nature Biotechnology*, **34**(2), 184.

Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. (1998). *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press.

Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics*, **14**(9), 755–763.

Forney, G. D. (1973). The viterbi algorithm. *Proceedings of the IEEE*, **61**(3), 268–278.

Fu, Y., Foden, J. A., Khayter, C., Maeder, M. L., Reyon, D., Joung, J. K., and Sander, J. D. (2013). High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nature Biotechnology*, **31**(9), 822–826.

Fusi, N., Smith, I., Doench, J., and Listgarten, J. (2015). In Silico Predictive Modeling of CRISPR/Cas9 guide efficiency. *bioRxiv*, page 021568.

Gagnon, J. A., Valen, E., Thyme, S. B., Huang, P., Ahkmetova, L., Pauli, A., Montague, T. G., Zimmerman, S., Richter, C., and Schier, A. F. (2014). Efficient mutagenesis by Cas9 protein-mediated oligonucleotide insertion and large-scale assessment of single-guide RNAs. *PLOS ONE*, **9**(5), e98186.

Haeussler, M., Schönig, K., Eckert, H., Eschstruth, A., Mianné, J., Renaud, J.-B., Schneider-Maunoury, S., Shkumatava, A., Teboul, L., Kent, J., *et al.* (2016). Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biology*, **17**(1), 148.

Hart, T., Chandrashekhar, M., Aregger, M., Steinhart, Z., Brown, K. R., MacLeod, G., Mis, M., Zimmermann, M., Fradet-Turcotte, A., Sun, S., *et al.* (2015). High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities. *Cell*, **163**(6), 1515–1526.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., *et al.* (2002). The Ensembl genome database project. *Nucleic Acids Research*, **30**(1), 38–41.

Huo, L., Zhang, H., Huo, X., Yang, Y., Li, X., and Yin, Y. (2017). pHMM-tree: phylogeny of profile hidden Markov models. *Bioinformatics*, page btw779.

Karplus, K., Barrett, C., and Hughey, R. (1998). Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, **14**(10), 846–856.

Kaur, K., Gupta, A. K., Rajput, A., and Kumar, M. (2016). ge-CRISPR-An integrated pipeline for the prediction and analysis of sgRNAs genome editing efficiency for CRISPR/Cas system. *Scientific Reports*, **6**.

Kim, D., Bae, S., Park, J., Kim, E., Kim, S., Yu, H. R., Hwang, J., Kim, J.-I., and Kim, J.-S. (2015). Digenome-seq: genome-wide profiling of CRISPR-Cas9 off-target effects in human cells. *Nature Methods*, **12**(3), 237–243.

Kleinstiver, B. P., Pattanayak, V., Prew, M. S., Tsai, S. Q., Nguyen, N. T., Zheng, Z., and Joung, J. K. (2016). High-fidelity CRISPR–Cas9 nucleases with no detectable genome-wide off-target effects. *Nature*, **529**(7587), 490–495.

Konermann, S., Brigham, M. D., Trevino, A. E., Joung, J., Abudayyeh, O. O., Barcena, C., Hsu, P. D., Habib, N., Gootenberg, J. S., Nishimasu, H., *et al.* (2015). Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature*, **517**(7536), 583.

Le Novere, N. (2001). MELTING, computing the melting temperature of nucleic acid duplex. *Bioinformatics*, **17**(12), 1226–1227.

Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American statistical Association*, **62**(318), 399–402.

Mali, P., Yang, L., Esvelt, K. M., Aach, J., Guell, M., DiCarlo, J. E., Norville, J. E., and Church, G. M. (2013). RNA-Guided Human Genome Engineering via Cas9. *Science*, **339**(6121), 823–826.

Mao, K. Z. (2004). Orthogonal forward selection and backward elimination algorithms for feature subset selection. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, **34**(1), 629–634.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, **405**(2), 442–451.

Men, K., Duan, X., He, Z., Yang, Y., Yao, S., and Wei, Y. (2017). CRISPR/Cas9-mediated correction of human genetic disease. *Science China Life Sciences*, pages 1–11.

Moreno-Mateos, M. A., Vejnar, C. E., Beaudoin, J.-D., Fernandez, J. P., Mis, E. K., Khokha, M. K., and Giraldez, A. J. (2015). CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. *Nature Methods*, **12**(10), 982–988.

Rahman, M. K. and Rahman, M. S. (2017). CRISPRpred: A flexible and efficient tool for sgRNAs on-target activity prediction in CRISPR/Cas9 systems. *PloS one*, **12**(8), e0181943.

Schliep, A., Schönhuth, A., and Steinhoff, C. (2003). Using hidden Markov models to analyze gene expression time course data. *Bioinformatics*, **19**(suppl 1), i255–i263.

Shalem, O., Sanjana, N. E., Hartenian, E., Shi, X., Scott, D. A., Mikkelsen, T. S., Heckl, D., Ebert, B. L., Root, D. E., Doench, J. G., *et al.* (2014). Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science*, **343**(6166), 84–87.

Shen, B., Zhang, W., Zhang, J., Zhou, J., Wang, J., Chen, L., Wang, L., Hodgkins, A., Iyer, V., Huang, X., *et al.* (2014). Efficient genome modification by CRISPR-Cas9 nickase with minimal off-target effects. *Nature Methods*, **11**(4), 399–402.

Swiech, L., Heidenreich, M., Banerjee, A., Habib, N., Li, Y., Trombetta, J., Sur, M., and Zhang, F. (2015). In vivo interrogation of gene function in the mammalian brain using CRISPR-Cas9. *Nature Biotechnology*, **33**(1), 102–106.

Torlay, L., Perrone-Bertolotti, M., Thomas, E., and Baciu, M. (2017). Machine learning–XGBoost analysis of language networks to classify patients with epilepsy. *Brain Informatics*, pages 1–11.

Tsai, S. Q., Zheng, Z., Nguyen, N. T., Liebers, M., Topkar, V. V., Thapar, V., Wyvekens, N., Khayter, C., Iafrate, A. J., Le, L. P., *et al.* (2015). GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nature Biotechnology*, **33**(2), 187–197.

Varshney, G. K., Pei, W., LaFave, M. C., Idol, J., Xu, L., Gallardo, V., Carrington, B., Bishop, K., Jones, M., Li, M., *et al.* (2015). High-throughput gene targeting and phenotyping in zebrafish using CRISPR/Cas9. *Genome Research*, **25**(7), 1030–1042.

Wang, T., Wei, J. J., Sabatini, D. M., and Lander, E. S. (2014). Genetic Screens in Human Cells Using the CRISPR-Cas9 System. *Science*, **343**(6166), 80–84.

Wheeler, T. J., Clements, J., Eddy, S. R., Hubley, R., Jones, T. A., Jurka, J., Smit, A. F., and Finn, R. D. (2013). Dfam: a database of repetitive DNA based on profile hidden Markov models. *Nucleic Acids Research*, **41**(D1), D70–D82.

Wong, N., Liu, W., and Wang, X. (2015). WU-CRISPR: characteristics of functional guide RNAs for the CRISPR/Cas9 system. *Genome Biology*, **16**(1), 218.

Xu, H., Xiao, T., Chen, C.-H., Li, W., Meyer, C. A., Wu, Q., Wu, D., Cong, L., Zhang, F., Liu, J. S., *et al.* (2015). Sequence determinants of improved CRISPR sgRNA design. *Genome Research*, **25**(8), 1147–1157.

Yin, C., Zhang, T., Qu, X., Zhang, Y., Putatunda, R., Xiao, X., Li, F., Xiao, W., Zhao, H., Dai, S., *et al.* (2017). In Vivo Excision of HIV-1 Provirus by saCas9 and Multiplex Single-Guide RNAs in Animal Models. *Molecular Therapy*, **25**(5), 1168–1186.

Yu, W., Mookherjee, S., Chaitankar, V., Hiriyanna, S., Kim, J.-W., Brooks, M., Ataeijannati, Y., Sun, X., Dong, L., Li, T., *et al.* (2017). Nrl knockdown by AAV-delivered CRISPR/Cas9 prevents retinal degeneration in mice. *Nature Communications*, **8**.

Zhang, L., Ai, H., Chen, W., Yin, Z., Hu, H., Zhu, J., Zhao, J., Zhao, Q., and Liu, H. (2017). CarcinoPred-EL: Novel models for predicting the carcinogenicity of chemicals using molecular fingerprints and ensemble learning methods. *Scientific Reports*, **7**(1), 2118.