Computational Biology and Chemistry 87 (2020) 107281

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/cbac

Research Article

SETE: Sequence-based Ensemble learning approach for TCR Epitope binding prediction

Yao Tong[a,c], Jiayin Wang[a,c,*], Tian Zheng[a,c], Xuanping Zhang[a,c], Xiao Xiao[a,c], Xiaoyan Zhu[a,c], Xin Lai[a,c], Xiang Liu[b,**]

[a] School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, 710049, China
[b] Department of Cardiothoracic Surgery, The Second Affiliated Hospital, University of South China, Hengyang, 421001, China
[c] Shaanxi Engineering Research Center of Medical and Health Big Data, Xi'an Jiaotong University, Xi'an, 710049, China

ARTICLE INFO

Keywords:
TCR
CDR3
VDJdb
Gradient boosting tree
Immunotherapy

ABSTRACT

Predicting the binding of T cell receptors (TCRs) to epitopes plays a vital role in the immunotherapy, because it guides the development of therapeutic vaccines and cancer treatments. Many prediction methods attempted to explain the relationship between TCR repertoires from different aspects such as the V(D)J gene locus and the biophysical features of amino acids molecules, but the extraction of these features is time consuming and the performance of these models are limited. Few studies have investigated how k-mers formed by adjacent amino acids in TCR sequences direct the epitope recognition, and the specific mechanism of TCR epitope binding is still unclear. Motivated by these, we presented *SETE* (Sequence-based Ensemble learning approach for TCR Epitope binding prediction), a novel model to predict the TCR epitope binding accurately. The model deconstructed the CDR3β sequence to short amino acid chains as features and learned the pattern of them between different TCR repertoires with gradient boosting decision tree algorithm. Experiments have demonstrated that *SETE* can be helpful in predicting the TCRs' corresponding epitopes and it outperforms other state-of-the-art methods in predicting the epitope specificity of TCR on VDJdb data set. The source codes have been uploaded at https://github.com/wonanut/SETE for academic usage only.

1. Introduction

To reactivate the immune system and fight against intruder in our body, immunotherapy tries to develop agents to repower the body's immune system (Sharma et al., 2011). Different from previous treatment strategies like physical resection, chemotherapy, radiotherapy and targeted therapy, the target of immunotherapy is not tumor cells or tissues, but the body's own immune system. Therefore, such therapeutic strategy fundamentally fights against the cancer and has a long-lasting response. T cells play an important role in the immunotherapy since they can recognize cancer cells using their specific immunity, and experiments have demonstrated that T cell receptor (TCR) genes isolated from antigen-specific T cells can be used as universal therapeutic molecules for antigen-specific immunotherapy (Xue et al., 2005). To date, numerous studies have attempted to explain how T cells recognize corresponding cancer cells so as to find a way to treat cancers (Kirsch et al., 2020), but the nature of the binding mechanism remains unclear.

In the process of recognizing T cell specificity, the TCR is required to be matched with the MHC molecule and its presented peptide (Rudolph et al., 2006). The diversity of MHC is limited for it recognizes peptides in a fixed manner (Lund et al., 2004). On the contrary, TCR gene fragments are obtained through a series of non-homologous recombination involving the combination of TCR loci from the variable region (V), diversity (D) and junction (J) of gene segments, and random nucleotides insertion and / or deletion (Bassing et al., 2002; Cabaniols et al., 2001), which produce a large number of different TCR sequences and the diversity estimated can reach $10^{15} \sim 10^{61}$ (Davis and Bjorkman, 1988; Mora and Walczak, 2016). Correspondingly, T cells in our body can specifically recognize a variety of case-specific antigens (epitopes). Furthermore, it has been observed that a TCR can specifically bind to multiple pMHCs because of the existence of cross-reactivity (Wooldridge et al., 2012), and the same pMHC can also bind to multiple TCRs (Sewell, 2012). On account of the multiplicity of relationships between TCR and epitopes, finding the mechanism by which TCR recognizes epitopes still remains a challenge (Miho et al., 2020).

In recent years, high-throughput sequencing has generated a large

* Corresponding author at: School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, 710049, China.
** Corresponding author.
E-mail addresses: wangjiayin@mail.xjtu.edu.cn (J. Wang), crisis163@163.com (X. Liu).

https://doi.org/10.1016/j.compbiolchem.2020.107281
Received 31 March 2020; Accepted 9 May 2020
Available online 20 June 2020

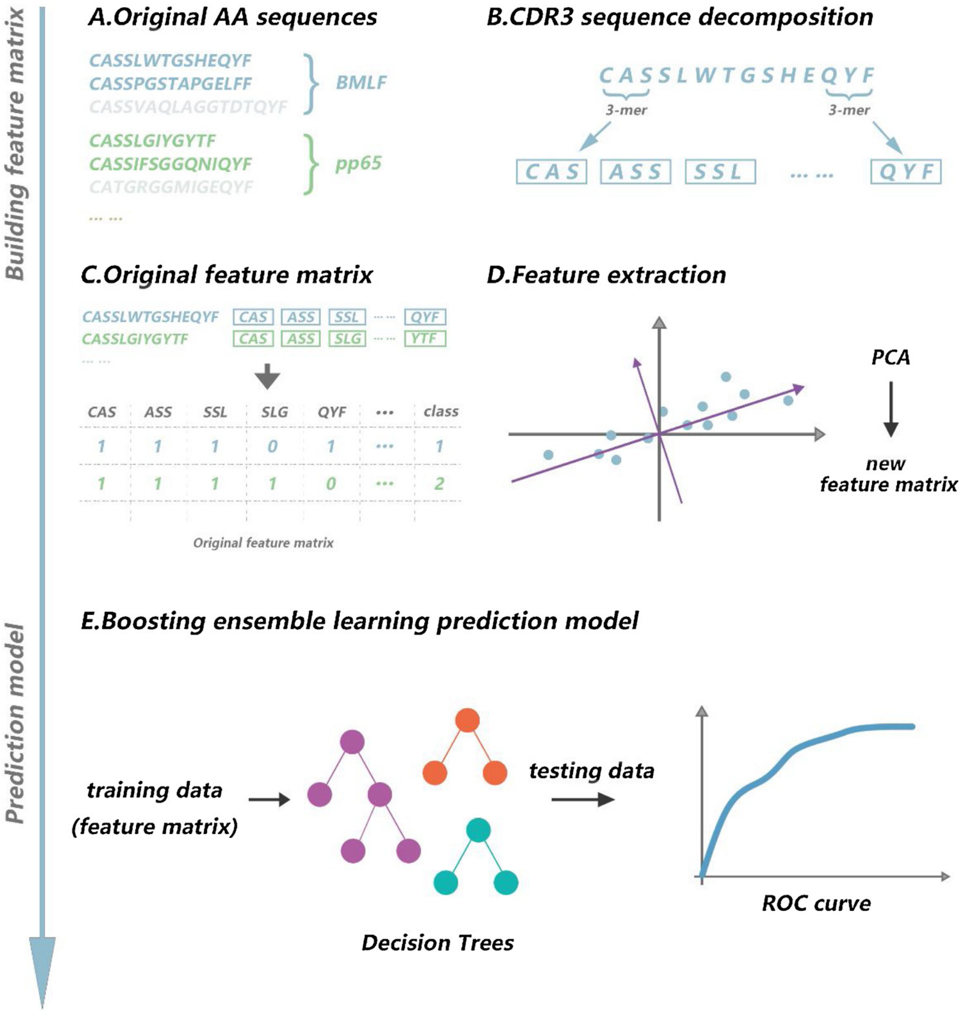

Fig. 1. The work flow of *SETE*. (A) The CDR3 sequences and their binding epitopes are collected as input. (B) Encode the CDR3 sequence into original feature matrix. Each sequence is deconstructed into overlapping contiguous short amino acid chains in length of 3. (C) All the deconstructed motifs form the original feature matrix. (D) Extract features and reduce the dimension with PCA to get the final feature matrix. (E) Feed the feature matrix to GBDT classifier to test the classification ability of *SETE*.

amount of available TCR data and annotation data (Mikhail et al., 2017; Tickotsky et al., 2017), which provide us with a chance to explore the relationship between TCR and epitopes based on machine learning frameworks. Researchers have attempted to predict TCRs' binding epitopes using a variety of machine learning methods and numerous computational methods have been proposed to solve the problem. The existing approaches can be divided into clustering tasks and classification tasks: (1) Rather than using the TCR and epitopes binding information, the clustering methods define some metrics to describe the similarity in two given TCR sequences and combine the two dimensional similarity distance between TCR sequences with clustering algorithms to evaluate the model performance (Meysman et al., 2018). However, limited by the amount of TCR and epitopes data, the known TCR sequence of a certain epitope remains rare and it is still a tough task for clustering algorithms to discover the hidden patterns under TCR sequences. (2) The classification methods try to find the most appropriate binding type between the TCR and the epitope using various classifier such as support vector machines, k nearest neighbor, random forests, convolutional neural networks, and so on. Glanville et al. developed a clustering-based method called *GLIPH* (Glanville et al., 2017), which attempted to cluster the TCRs by both local similarity and global similarity, and then the TCR sequences satisfying certain threshold among the clustering network will be clustered into specific repertoires. However, since some original information between TCR sequences was lost in the process of clustering, only a small percentage of data could be clustered and a large number of sequences were regarded as outliers. The *TCRdist* proposed by Dash et al. (2017) used the BLOSUM62 matrix to calculate the distance between two TCR sequences, and then used the k nearest neighbor classifier to predict whether a given TCR is similar to some certain TCR clusters. Nevertheless, encoding amino acid using BLOSUM has two disadvantages. Firstly, because the length of the TCR chain is not fixed, the TCR chain must be aligned before encoding. Although the CDR3s can be aligned by introducing a gap in the middle of the amino acid sequence (Jokinen et al., 2019; Lefranc et al., 2005), it will change the original sequence anyway. Secondly, the BLOSUM matrix coding method exploits a priori knowledge of similarities and dissimilarities between amino acids (Nielsen et al., 2003), which makes the results less convincing. De Neuter et al. (2018) extracted the biophysical properties of TCR amino acid sequences, such as helicity and hydrophobicity, as features of TCR sequences, and then used the random forest to predict epitopes recognized by TCRs. However, none of them fully utilize the adjacent relationship between the amino acids on the TCR chain.

Furthermore, previous studies have shown that different regions in the TCR sequences have different contribution to the specific recognition of TCR (Calis et al., 2013). The most variable region is CDR3, which is most likely to determine whether the TCRs can interact with the epitopes (Jorgensen et al., 1992), and the β-chain of this region is the major determinant of the recognition of peptides (Glanville et al., 2017).

Motivated by these, we present *SETE*, a novel computational approach for predicting epitope by combining a sequence-based feature encoding method and gradient boosting decision tree (Friedman, 2001) (GBDT) classifier. Firstly, we focus on the effect of adjacent amino acids in the CDR3 region on the specific recognition of TCR in this paper. The β-chain of CDR3 sequences are deconstructed into overlapping contiguous short amino acid chains and we extracted the statistical result of k-mer as the original feature. Secondly, due to the local similarity between TCR sequence, we perform principal component analysis (PCA) on the initial feature matrix to remove redundant information and reduce the dimension of the feature matrix. Thirdly, we input the previously extracted feature matrix to the gradient boosting decision tree classifier and output the predict epitopes of TCR sequences. In order to evaluate the performance of *SETE*, we test it on data provided by Dash et al. and public data set VDJdb. The multi-class results showed that *SETE* outperforms other methods in predicting the epitope of TCR and the adjacent relationship of amino acids in the CDR3β chain can facilitate TCR recognition the TCR and help identify the corresponding epitope specifically.

2. Methods

The workflow of *SETE* is shown in Fig. 1, where the input is a set of CDR3 sequences with corresponding epitopes label as category labels. First, we remove the TCR-epitope pairs which do not meet the minimum limitation because of the high diversity of TCR sequences, only if the CDR3 sequences in certain repertoire are abundant enough can they be meaningful to represent their corresponding repertoire. Therefore, it is necessary to delete the TCR-epitope pairs which do not meet the minimum limitation. Secondly, we encode the original sequences data because the length of CDR3β is not fixed. A feature matrix will be built for the filtered data and it will be deconstructed into overlapping contiguous short amino acids chains and the statistical result of the k-mers will form the initial feature matrix. Thirdly, due to the high dimension of the original feature matrix and the local similarity between TCR sequences, the PCA is applied on the initial feature matrix to reduce the dimension of the feature matrix. Finally, the feature matrix will be sent into the model for training, and a classification decision model consisting of several classification regression decision trees is obtained.

2.1. Sequence encoding

There are several methods to encode the raw TCR sequence before it can be input to the model: 1) the conventional sparse coding encodes each amino acid into a 20-digit binary number (one 1 and nineteen 0), which produces a very sparse feature matrix. 2) the BLOSUM50 or BLOSUM62 (Henikoff ;, 1992) matrix encoding scheme encodes the amino acid as the fraction of the corresponding column of the amino acid in the substitution matrix. 3) the physicochemical characteristics of amino acids can also be used to encode TCR sequences (Zhou et al., 2017). Nevertheless, some encoding methods mentioned above require alignment of the TCR referring to the IMGT standard (Lefranc et al., 2005) for the TCR sequences are amino acid sequences of unequal lengths, while others need additional information. Is there any other encoding approach? A previous studyshowed that the statistical results of short amino acids motifs (with lengths of 2, 3 and 4) in the high-contact-probability region of CDR3 are useful for identifying diff ;erent types of TCRs (Glanville et al., 2017). In order to avoid the alignment operation, we encode the CDR3β sequence by cutting it into short chains with length of 3 amino acids by referring to the method of Cinelli et al. (2017). So each CDR3 sequence will be represented by several short stretches of contiguous amino acids.

Fig. 2 is a feature matrix extracted from diff ;erent types of TCR sequences and features were selected using chi-square test. In Fig. 2(a), 200 features were remained and in Fig. 2(b), only 20 features were remained. It was obvious that in some epitopes the feature patterns were similar, such as epitope gene M1 and PB1. We expected to find the hidden pattern of the feature matrix through suitable machine learning algorithms that can help discover how the TCRs recognize epitopes. Since most of the CDR3β chains start with a' CAS' amino acids motif and end with' QYF', there are two thin lines that are common to the feature maps of the individual epitopes. In addition, the feature maps of diff ;erent epitopes have obvious diff ;erences, while the features of the same epitope have obvious similarities.

2.2. Features extraction

There are two reasons to conduct feature extraction: First, there are 20 kinds of amino acids in total, the short-chain of amino acids with a length of 3 will have up to 20^3 kinds of combinations and the feature can reach up to 8000 dimensions, so it is necessary to reduce the dimension; Second, due to the global similarity between similar TCR sequences, there may be redundant information between the 3-mers of the same class of TCR sequences. Here we use PCA to reduce the dimension of the original feature matrix. Supposed that the original feature matrix is $\boldsymbol{X} = \{x_1, x_2, ..., x_n\}$, with each $x_\mathrm{i} \in \mathbb{R}^m$ viewed as a column vector, n represents the number of training samples and m represents the dimension of features. Firstly, each column of feature needs to be centralized:

$$x_\mathrm{i} = x_i - \frac{1}{n}\sum_{i=1}^{n} x_i$$

Assuming that the projection of x_i on the hyperplane in the new space is $\boldsymbol{W}^T x_i$, the variance between x_i after projection should be maximized if all the samples are separated as much as possible,. The optimization target is:

$$\max_{W} tr(\boldsymbol{W}^T \boldsymbol{X}\boldsymbol{X}^T \boldsymbol{W})\ s.\ t.\ \ \boldsymbol{W}^T\boldsymbol{W} = I$$

Solving the above equation with Lagrangian multiplier method, we will get $\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{W} = \lambda \boldsymbol{W}$. Therefore, the problem will be solved if we decompose the covariance matrix $\boldsymbol{X}\boldsymbol{X}^T$, and sort the obtained eigenvalues from large to small: $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$. Then take the eigenvectors corresponding to the first k eigenvalues to form a projection matrix $\mathbf{W} = (\mathrm{w}_1, w_2, ..., w_k)$, the final feature matrix $\boldsymbol{W}^T\boldsymbol{X}$ is a matrix of k rows and n columns.

2.3. Gradient boosting decision tree model

As an ensemble learning algorithm using boosting strategy, Gradient Boosting Decision Tree (GBDT) classifier is widely used for its powerful classification and regression performance. Different from other ensemble techniques like Adaboost and random forest, the basic idea of GBDT is to train the newly added weak classifier according to the negative gradient information of the current model loss function, and then combine the trained weak classifiers into the existing model in an accumulated form. In each iteration, the negative gradient of the current model on all samples is first calculated, and then a new weak classifier is trained based on the negative gradient information. On the one hand, GBDT has great generalization ability, and it can automatically find the high-order relationship between features. On the other hand, GBDT builds a prediction model in a stage-wise fashion to reduce the bias of the model constantly. Therefore, it is suitable for handling the TCR classification tasks.

Supposed that we are given n training data $\{(x_1, y_1), ..., (x_n, y_n)\}$, where $x_i \in \boldsymbol{X} \subseteq \mathbb{R}^n$, $y_i \in \boldsymbol{Y} \subseteq \mathbb{R}$, $i = 1,2, ..., n$. After inputting the prediction data x, the gradient boosting decision tree makes predictions by linearly combining the decision results of all the trees:

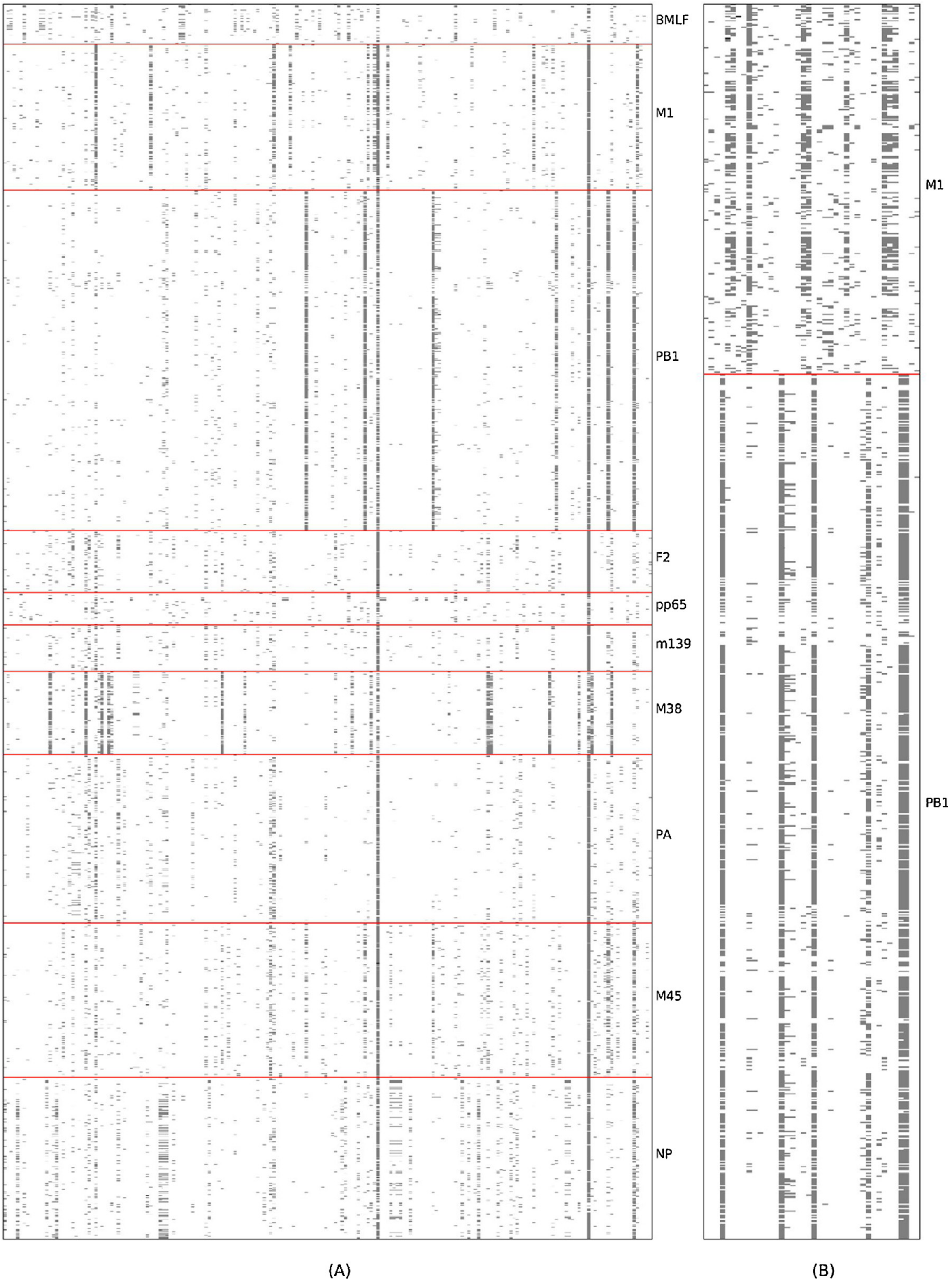

Fig. 2. TCRs binding to diff ;erent epitopes show diff ;erent patterns. Each row corresponding to a single CDR3 sequence and each column represents a unique motif. (A) shows the feature patterns among different epitopes, 200 features were selected. (B) shows the feature patterns of M1 and PB1, 40 features were selected.

$$f_M(x) = \sum_{m=1}^{M} \beta_m T(x; \Theta_m)$$

where $T(x; \Theta_m)$ is the m-th decision tree and β_{m} is the weight of the tree, Θ_m is the parameter of the tree, M represents the number of decision tree (the iterations to train the model).

The gradient boosting decision tree constructs the model in a forward-accumulated manner, a new decision tree will be added to the classification for each iteration. Assume that the initial model is:

$$f_0(x) = arg\min_{c} \sum_{i=1}^{N} L(y_i, c)$$

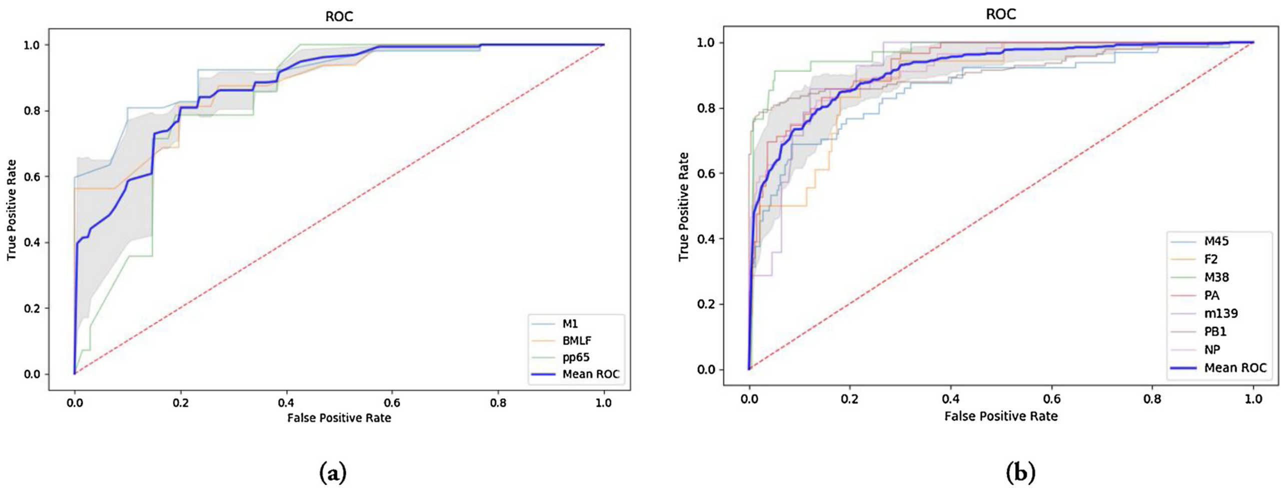

Fig. 3. ROC curves for each TCR repertoires in the Dash et al. data set. Each translucent curve represents the ROC curve of a repertoires. The blue curve is the mean ROC curve of the ten repertoires. The translucent area colored by gray shows then standard deviation of the TPR. (a) ROC curves for 3 human TCR repertoires. The mean AUC is 0.849. (b) ROC curves for 7 mouse TCR repertoires. The mean AUC is 0.897.

where function L represent a log-likelihood loss function (deviance), which is defined as:

$$L(\boldsymbol{Y}, P(\boldsymbol{YX}) = -logP(\boldsymbol{YX}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(p_{ij}))$$

where $\boldsymbol{Y}$ is the output variable, $\boldsymbol{X}$ is the input variable, M is the number or class of epitopes. y_{ij} is a binary indicator that representing whether the category j is the true category of the input data x_i. p_{ij} is the probability that x_i belongs to category j predicted by classifier.

Then the process of the m-th iteration of the model is:

$$f_m(x) = f_{m-1}(x) + \beta_m T(x; \Theta_m)$$

where $f_{m-1}(x)$ is the decision model of the $(m-1)$th iteration. After the m-th iteration, a new decision tree will be added to $f_{m-1}(x)$.

The parameter of decision tree $T(x; \Theta_m)$ is learned following this principle:

$$\hat{\Theta}_m = \underset{\Theta_m}{\operatorname{argmin}}\sum_{i=1}^{N} L(y_i, f_{m-1}(x_i) + \beta_m T(x; \Theta_m))$$

Since the meta classifiers are linearly added, the goal is to estimate the parameter Θ_m using the residual $L(y, f_{m-1}(x))$. To this end, the negative gradient of the loss function on the model f_{m-1} can be used to estimate the residual:

$$R_{\mathrm{mi}} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{m-1}(x)}$$

Where i is the index of the i-th training sample. Finally, all the collections of R_{mi} will be used to fit a regression classification decision tree (CART) to calculate the parameter Θ_m. The GBDT algorithm is implemented using scikit-learn (Pedregosa et al., 2020).

3. Results

To evaluate the performance of the proposed method, we first examined the performance of *SETE* in predicting epitopes on dataset provided by Dash et al. Then, we compared the classification performance of *SETE* with the existing state-of-the-art prediction models *TCRdist* and *TCRGP* (Jokinen et al., 2019) on public dataset VDJdb. Since we focused on the β-chain of CDR3 sequence, only sequences in this areawere used to develop the feature matrix.

3.1. Data set

The data set provided by Dash et al. contains 2336 annotated TCR sequences from 10 different epitopes, including three kinds of human epitopes and seven mouse epitopes. To fully explore the determinants of epitope speficity considering the difference between different species, we also train the models on mouse and human datasets respectively. All the three kinds of human epitopes (GILGFVFTL, GLCTLVAML and NLVPMVATV) are presented by HLA-A*02:01, while the mouse epitopes are presented by D^b (LSLRNPILV, ASNENMETM, SSLENFRAYV and HGIRNASFI) and K^b (SSYRRPVGI, TVYGFCLL and SSPPMFRV). The quantity of TCR repertoires that bind different epitopes varies greatly. For instance, there are only 61 TCR sequences of the human epitope gene pp65, while the mouse epitope gene PB1 has 642 TCR sequences data. To ensure the generalization of the model, we use the receiver operating characteristic (ROC) curve to evaluate *SETE* under five-fold cross-validation. For each of the five subsets, the training set and test set is randomly partitioned.

3.2. Performance of the proposed method

Since Gradient Boosting Decision Tree is an ensemble learning algorithm based on Decision Tree algorithm, most of the parameters are the same as that on Decision Tree algorithm. We search the best parameters of GBDT with the help of grid search algorithm. The following parameters are considered: iteration number (n_estimators), learning rate (learning_rate), max depth of the tree (max_depth) and the number of features to consider when looking for the best split (max_features). Finally, the best parameters for Dash et al. data set we found are n_estimators = 70, learning_rate = 0.1, max_depth = 10 and max_features = 'sqrt'.

Firstly, We evaluated the multi-class performance of *SETE* without extracting features from the original features matrix. The ROC curves for each epitope in the human and mouse were presented in Fig. 3. The average AUC was 0.849 for human dataset and 0.897 for mouse dataset. The area under receiver operating characteristic curves (AUROC) was greater than 0.8 for all epitopes except pp65, which might be due to the limited quantity and the high diversity of TCR sequences in that repertoires.

In addition, we tested the prediction power of *SETE* using the feature extraction method metioned before and the detailed result was shown in Table 1. From the table we can see that although all the repertoires have auc score greater than 0.8, the final result varies notably between repertoires. It is clearly that repertoires with larger sequence

Table 1
Five-fold cross-validation on human and mouse dataset.

Species	Epitope gene	Sequence number	Cross-validation results (AUC)					
			Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Avg. ± Std.
Human	pp65	61	0.864	0.804	0.731	0.831	0.801	0.806 ± 0.044
	BMLF	76	0.973	0.890	0.877	0.881	0.938	0.912 ± 0.038
	M1	275	0.918	0.918	0.913	0.865	0.903	0.903 ± 0.020
Mouse	M139	87	0.815	0.905	0.798	0.860	0.820	0.840 ± 0.038
	PA	324	0.933	0.938	0.924	0.955	0.928	0.936 ± 0.011
	M38	158	0.951	0.926	0.944	0.980	0.965	0.953 ± 0.018
	M45	291	0.895	0.894	0.850	0.909	0.946	0.899 ± 0.031
	PB1	642	0.940	0.932	0.942	0.954	0.918	0.937 ± 0.012
	NP	305	0.924	0.941	0.952	0.922	0.963	0.940 ± 0.016
	F2	117	0.891	0.847	0.877	0.855	0.870	0.868 ± 0.016

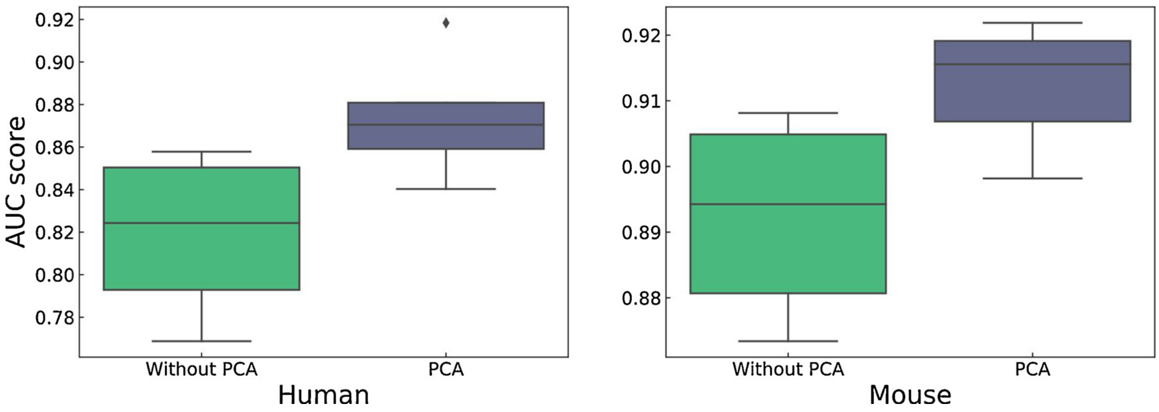

Fig. 4. Comparision of the impact of feature extraction on *SETE*.The left panel are comparative boxplot for human and the right one for mouse.

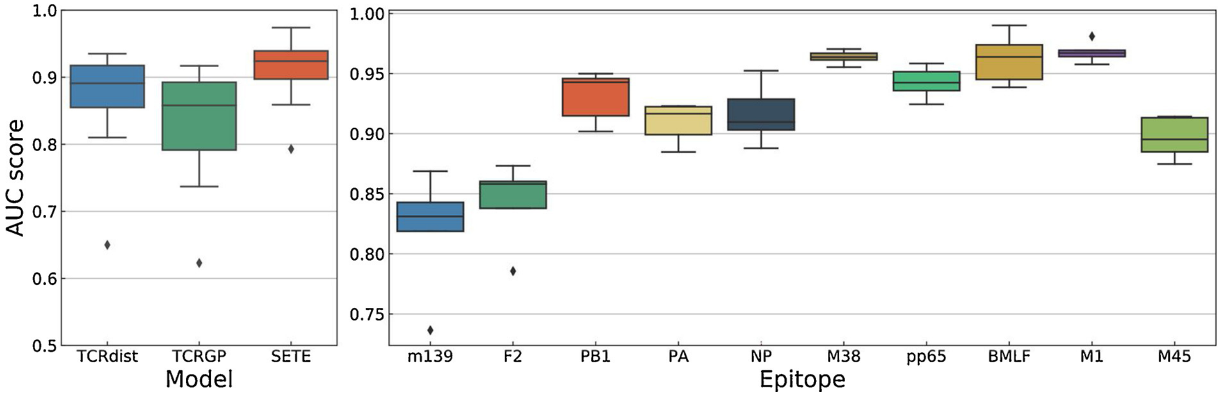

Fig. 5. Five-fold cross-validation results on Dash data set. The left panel shows the AUROC-scores of different models. The results of *TCRdist* were obtained from the original paper and the results of *TCRGP* were collected under the default parameters. The results of *SETE* were multi-class AUROC-scores, for each kind of epitope we took the mean five-fold cross-validation AUC score. The right panel shows the multi-class AUROC-scores for each kind of epitope.

number tend to have higher AUC score, so the differences may be explained by the limited number of available training data. What's more, Dash et al. has metioned that pp65 is the most diverse repertoire, so the high diversity is another reason why the pp65 repertoire is indistinguishable.

3.3. *Evaluating the significance of PCA*

In order to test the effect of PCA on the performance of *SETE*, we compared the classification efficiency of extracted features tothat of the original features. The results indicated that the performance of our model is better and more stable on both the human and mouse data set after feature extraction (Figure 4). The AUC score was raised at least by 4 % and 2 % respectively on the two data sets.

The comparative experiments proved that the PCA process was capable of reducing the dimension of the original features and improving the performance of *SETE*.

3.4. *Comparison between SETE and other models*

So far, several TCR prediction models have been proposed. Noting that the recently proposed methods *TCRGP* was tested on data sets provided by Dash and VDJdb, in order to highlight the performance of our model, we compared *SETE* with the state-of-the-art methods on the two data sets, and used the ROC as the criterion.

Firstly, we compared the performance of *SETE* with *TCRdist* and *TCRGP* on Dash's dataset. Since the *TCRdist* was designed for all the regions on TCR sequences (CDR1, CDR2, CDR2.5 and CDR3), but *SETE*

Table 2
Five-fold cross-validation on VDJdb dataset.

Epitope	Sequence number	Cross-validation results (AUC)					
		Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Avg. + Std.
GLCTLVAML	299	0.92	0.93	0.94	0.95	0.92	0.93 ± 0.01
FLKEKGGL	104	0.84	0.86	0.88	0.82	0.85	0.84 ± 0.01
PKYVKQNTLKLAT	56	0.60	0.69	0.68	0.58	0.75	0.66 ± 0.06
RPRGEVRFL	68	0.99	0.90	0.88	0.98	0.94	0.93 ± 0.04
KAFSPEVIPMF	134	0.85	0.89	0.87	0.90	0.81	0.86 ± 0.03
TPQDLNTML	52	0.86	0.81	0.81	0.71	0.80	0.80 ± 0.04
YVLDHLIVV	66	0.81	0.79	0.73	0.86	0.86	0.81 ± 0.04
GPGHKARVL	62	0.77	0.76	0.73	0.66	0.71	0.72 ± 0.04
FRDYVDR—QASQE	141	0.87	0.95	0.99	0.98	0.93	0.94 ± 0.04
GTSGSPIVNR	51	0.59	0.78	0.70	0.67	0.88	0.72 ± 0.09
GTSGSPIINR	65	0.75	0.64	0.59	0.65	0.62	0.65 ± 0.05
GILGFVFTL	239	0.87	0.89	0.90	0.81	0.89	0.87 ± 0.03
KLVALGINAV	65	0.59	0.52	0.57	0.54	0.50	0.54 ± 0.03
IPSINVHHY	65	0.84	0.83	0.89	0.82	0.88	0.85 ± 0.02
NLVPMVATV	413	0.89	0.91	0.93	0.92	0.90	0.90 ± 0.01
KRWIILGLNK	212	0.89	0.86	0.82	0.89	0.83	0.85 ± 0.02
CINGVCWTV	76	0.91	0.90	0.96	0.78	0.88	0.88 ± 0.05
RAKFKQLL	225	0.84	0.90	0.90	0.90	0.92	0.89 ± 0.02
TPRVTGGGAM	184	0.94	0.93	0.88	0.91	0.92	0.91 ± 0.02
ATDALMTGY	152	0.84	0.82	0.90	0.93	0.90	0.87 ± 0.04
LLWNGPMAV	223	0.82	0.88	0.77	0.86	0.81	0.82 ± 0.03
EIYKRWII	81	0.89	0.75	0.78	0.85	0.86	0.82 ± 0.05

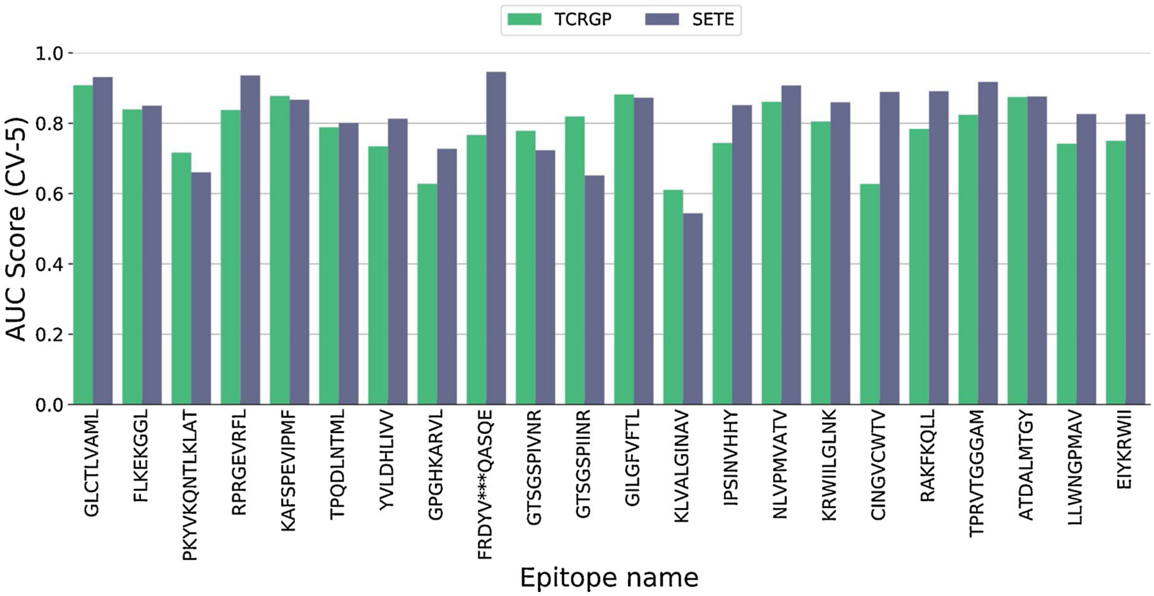

Fig. 6. Performance comparison between *SETE* and *TCRGP* on VDJdb data set.

Table 3
Five-fold cross-validation on Dash. dataset with duplicated sequences removed.

Species	Epitope gene	Sequence number	Cross-validation results (AUC)					
			Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Avg. ± Std.
Human	pp65	56	0.587	0.697	0.750	0.684	0.700	0.684 ± 0.053
	BMLF	55	0.806	0.848	0.829	0.905	0.820	0.842 ± 0.035
	M1	167	0.758	0.793	0.828	0.848	0.782	0.802 ± 0.032
Mouse	M139	73	0.738	0.709	0.748	0.750	0.693	0.728 ± 0.023
	PA	232	0.884	0.877	0.841	0.862	0.774	0.848 ± 0.040
	M38	65	0.757	0.716	0.836	0.832	0.809	0.788 ± 0.045
	M45	206	0.843	0.822	0.707	0.864	0.784	0.804 ± 0.055
	PB1	408	0.830	0.867	0.845	0.921	0.859	0.864 ± 0.031
	NP	151	0.731	0.802	0.801	0.827	0.814	0.795 ± 0.033
	F2	103	0.729	0.811	0.817	0.871	0.776	0.801 ± 0.047

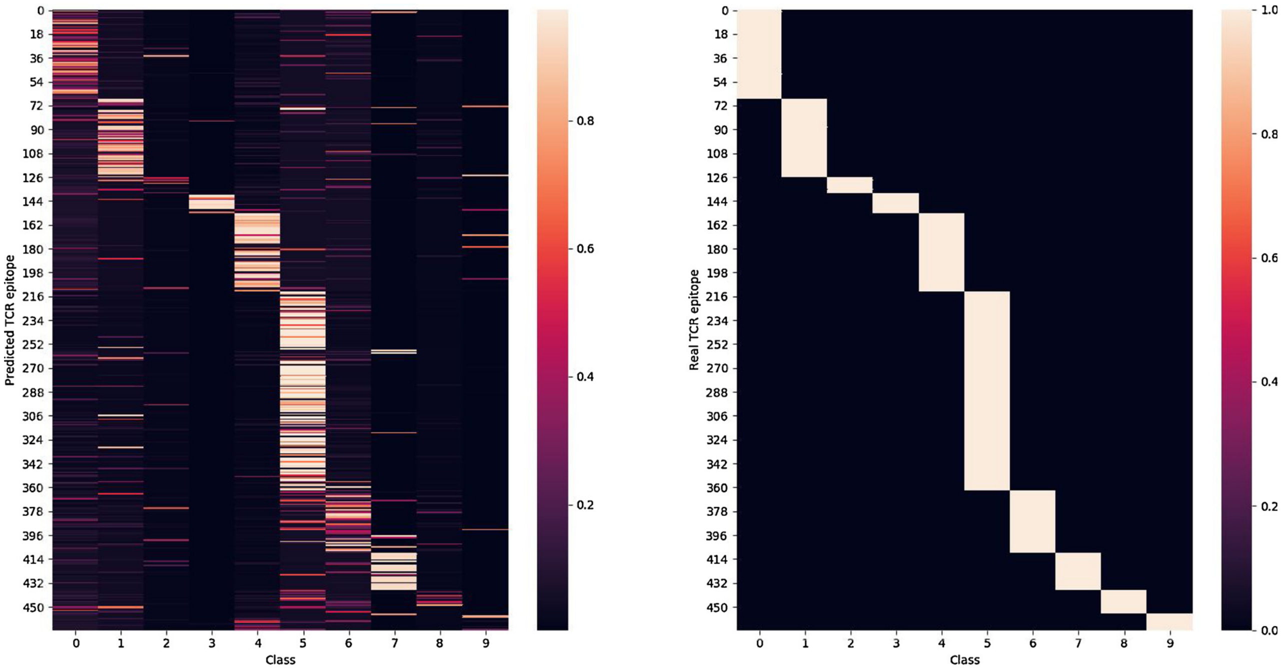

Fig. 7. The left panel are the probability distribution provided by the classifier. Each row represents a test sequence, and each column represents a epitopes. The brighter the region indicates that the higher the probability that a given TCR sequence belongs to the epitope. The right panel shows the one-hot encoding of the real label.

just aimed at the CDR3β sequence, we just compared *SETE* with the relevant results in the source paper. Similarly, the *TCRGP* can utilize the CDR3α and β sequence for prediction, and only the parts involving the CDR3β sequence were adopted for comparison. To make sure different models were compared in the same situation, we mixed the human and mouse epitopes data together as training set. We also used the five-fold cross validation to evaluate the performance. The distributions of mean AUC score for each repertoire were shown in Fig. 5. *SETE* outperformed *TCRdist* although it only utilized the CDR3β sequence for prediction. Moreover, *SETE* was more stable than the other two prediction models.

In addition, we compared *SETE* to *TCRGP* on VDJdb data set. VDJdb is a public database that collects annotated TCR sequences from published papers. We selected the TCR repertoires with sequence numbers greater than 50, and finally 22 repertoires were selected. Here we only compare *SETE* with *TCRGP* using β-chain of CDR3 sequence, and the result of *TCRGP* was collected under default parameters.

The cross-validation results as shown in Table 2 indicated that our classifier had ideal performance greatly except for several epitopes (PKYVKQNTLKLAT, GTSGSPIINR and KLVALGINAV) whose AUC scores were less than 0.7. The same problem can also be found in *TCRGP*, which might be caused by the limited amount of training data as discussed before.

Then we compared *SETE* with *TCRGP* as shown in Fig. 6 and *SETE* outperformed *TCRGP* in most cases. *SETE* had an average AUC score of 0.826 over the 22 classes of epitopes, higher than that on *TCRGP* (0.782). In several epitopes like KLVALGINAV, both *TCRGP* and *SETE* performed poorly, possibly because the lack of training data makes it hard for the classifier to distinguish TCR sequences binding to those epitopes.

The experiment results showed that the CDR3β chain played an important role in the prediction of TCRs epitope and the proposed model had great prediction performance. All the comparison experiments demonstrate that our model can outperform other existed TCR prediction methods in some cases and it might be a useful tool in the field of immunotherapy.

4. Discussion

4.1. Discussion about the duplicated sequences

We have to point out that there have some duplicated CDR3 sequences both in the clone sequences provided by Dash and VDJdb datasets which might be caused by the following reasons: (a) different individuals may share the same TCRs which are defined as public TCRs (DeWitt et al., 2018); (b) different nucleotide sequences may be translated into the same amino acid sequence (Yuval et al., 2018); (c) after the V(D)J combination (Bassing et al., 2002), TCR sequences can have different CDR3β but share the same CDR3α sequence. Moreover, the current datasets are collected from different published assays (Mikhail et al., 2017), so data from different assays might contains the same sequences. Here we follow the definition of unique TCR from *TCRGP* (Jokinen et al., 2019) that a TCR is unique if it has a unique combination of CDR3β and Vβ-gene. To explore the prediction ability of *SETE*, we did the same experiment on the dataset from which the duplicated sequences were removed and the results are shown in Table 3.

4.2. More discussion

Because how the TCR recognize corresponding epitope still remains unclear and such predictive models are urgently needed in the field of immunotherapy, the exising predictive models try to solve the problem using different strategies. But up to now, the performance of these models is not good enough, so it is necessary to develop new methods that take account of other features. Since prior studies have demonstrated the importance of CDR3 in distinguishing epitopes, we tried to build models using only CDR3 sequences data. In this paper, we tried to predict the epitopes using short amino acid chains cut from TCR sequences with Gradient Boosting Decision Tree. And the experiment results showed that *SETE* had great predictive ability. But due to the limited data, it is hard for *SETE* to distinguish different TCR sequences which are diverse.

To make up for this weakness, *SETE* can output the probability that the given sequence belongs to different epitopes, thus the prediction results can be used to narrow the search field. As is shown in Fig. 7, although the classifier might make mistakes in some situations, the given probability can help troubleshoot some wrong answers. Besides, because the high-quality TCRs with corresponding epitopes label are still rare, *SETE* cannot learn more feature patterns of TCR sequences and epitopes from limited data. We expect more available data in the future and check the classification eff ;ect of *SETE* in a larger training set with the development of high throughput sequencing technology.

5. Conclusion

The specific recognition of TCR-corresponding epitopes has important application value in immunotherapy. Existing prediction models attempt to encode TCR sequences using BLOSUM matrix, or encode them using biophysical properties of amino acids, and combined with the corresponding prediction model to predict the TCR epitope. However, the effect of k-mer on the epitope prediction of adjacent amino acids in the TCR sequence was not considered. In this paper, we proposed a prediction model using the 3-mer statistics of the TCR sequence as the initial feature matrix. We used the principal component analysis method to reduce the dimension of the feature matrix, and combined the gradient boosting decision tree ensemble learning model to predict the TCR-corresponding epitope. The experiment results showed that *SETE* can serve as a reliable tool in predicting possible epitopes. In addition, *SETE* can directly output the probability that a given TCR belongs to each epitope. In practical applications, the probability result can be used to further narrow the search field of TCR-corresponding epitopes, and to find T cells needed to be specifically identified.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Bassing, C.H., Swat, W., Alt, F.W., 2002. The Mechanism and Regulation of Chromosomal V(d)j Recombination. 11. .

Cabaniols, J.P., Fazilleau, N., Casrouge, A., Kourilsky, P., Kanellopoulos, J.M., 2001. Most α/β t cell receptor diversity is due to terminal deoxynucleotidyl transferase. J. Exp. Med. 194 (9), 1385–1390.

Calis, J.J.A., Maybeno, M., Greenbaum, J.A., Weiskopf, D., De Silva, A.D., Sette, A., Kemir, C., Peters, B., 2013. Properties of mhc class i presented peptides that enhance immunogenicity. PLoS Comput. Biol. 9 (10), 1–13.

Cinelli, M., Sun, Y., Best, K., Heather, J.M., Reich-Zeliger, S., Shifrut, E., Friedman, N., Shawe-Taylor, J., Chain, B., 2017. Feature selection using a one dimensional naive bayes classifier increases the accuracy of support vector machine classification of cdr3 repertoires. Bioinformatics 771.

Dash, P., Fiore-Gartland, A.J., Hertz, T., Wang, G.C., Sharma, S., Souquette, A., Crawford, J.C., Clemens, E.B., Nguyen, T.H.O., Kedzierska, K., 2017. Quantifiable predictive features define epitope-specific T cell receptor repertoires. Nature 547 (7661), 89–93.

Davis, M.M., Bjorkman, P.J., 1988. Erratum: T-cell antigen receptor genes and t-cell recognition. Nature 335 (6192), 744.

DeWitt, I., William, S., Smith, A., Schoch, G., Hansen, J.A., Matsen, I., Frederick, A., Bradley, P., 2018. Human t cell receptor occurrence patterns encode immune history, genetic background, and receptor specificity. eLife 7, e38358.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 29 (5), 1189–1232.

Glanville, J., Huang, H., Nau, A., Hatton, O., Wagar, L.E., Rubelt, F., Ji, X., Han, A., Krams, S.M., Pettus, C., 2017. Identifying specificity groups in the T cell receptor repertoire. Nature 547 (7661), 94–98.

Henikoff ;, H.J.G., 1992. Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. U. S. A. 89 (22), 10915–10919.

Jokinen, E., Huuhtanen, J., Mustjoki, S., Heinonen, M., L¨ahdesma¨ki, H., 2019. Determining epitope specificity of t cell receptors with tcrgp. bioRxiv 02.

Jorgensen, J.L., Esser, U., Fazekas de St. Groth, B., Reay, P.A., Davis, M.M., 1992. Mapping t-cell receptorpeptide contacts by variant peptide immunization of single-chain transgenics. Nature 355 (6357), 224–230.

Lefranc, M.P., Pommi, C., Kaas, Q., Duprat, E., Bosc, N., Guiraudou, D., Jean, C., Ruiz, M., Pidade, I.D., Rouard, M., 2005. Imgt unique numbering for immunoglobulin and t cell receptor constant domains and ig superfamily c-like domains. Dev. Comp. Immunol. 29 (3), 0–203.

Lund, O., Nielsen, M., Kesmir, C., Petersen, A., Lundegaard, C., Worning, P., Sylvester-Hvid, C., Lamberth, K., Roder, G., Justesen, S., 2004. Definition of supertypes for hla molecules using clustering of specificity matrices. Immunogenetics 55 (12), 797–810.

Meysman, P., De Neuter, N., Gielis, S., Bui Thi, D., Ogunjimi, B., Laukens, K., 2018. On the viability of unsupervised T-cell receptor sequence clustering for epitope preference. Bioinformatics 35 (9), 1461–1468.

Miho, E., Yermanos, A., Weber, C.R., Berger, C.T., Reddy, S.T., Greiff, V., 2018. Computational strategies for dissecting the high-dimensional complexity of adaptive immune repertoires 9, 224.

Mikhail, S., Bagaev, D.V., Zvyagin, I.V., Vroomans, R.M., Chase, C.J., Garry, D., Komech, E.A., Sycheva, A.L., Koneva, A.E., Egorov, E.S., 2017. Vdjdb: a curated database of t-cell receptor sequences with known antigen specificity. Nucleic Acids Res. (D1), D1.

Mora, T., Walczak, A., 2016. Quantifying Lymphocyte Receptor Diversity. 04. .

Neuter, N.D., Bittremieux, W., Beirnaert, C., Cuypers, B., Mrzic, A., Moris, P., Suls, A., Tendeloo, V.V., Ogunjimi, B., Laukens, K., Meysman, P., 2018. On the feasibility of mining cd8+ t cell receptor patterns underlying immunogenic peptide recognition. Immunogenetics 70 (3), 159–168.

Nielsen, M., Lundegaard, C., Worning, P., Lauemoller, S.L., Lamberth, K., Buus, S., Brunak, S., Lund, O., 2003. Reliable prediction of t-cell epitopes using neural networks with novel sequence representations. Protein Sci. 12 (5), 1007–1017.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Mller, A., Nothman, J., a. Louppe, G., 2012. Scikit-learn: Machine learning in python. Journal of Machine Learning Research 12 (10), 2825–2830.

Rudolph, M.G., Stanfield, R.L., Wilson, I.A., 2006. How tcrs bind mhcs, peptides, and coreceptors. Annu. Rev. Immunol. 24 (1), 419–466.

Sewell, A.K., 2012. Why must t cells be cross-reactive? Nat. Rev. Immunol. 12 (9), 669–677.

Sharma, P., Wagner, K., Wolchok, J.D., Allison, J.P., 2011. Novel cancer immunotherapy agents with survival benefit: recent successes and next steps. Nat. Rev. Cancer 11 (11), 805–812.

Tickotsky, N., Sagiv, T., Prilusky, J., Shifrut, E., Friedman, N., 2017. Mcpas-tcr: a manually-curated catalogue of pathology-associated t cell receptor sequences. Bioinformatics 33 05.

Wooldridge, L., Ekeruche-Makinde, J., van den Berg, H.A., Skowera, A., Miles, J.J., Tan, M.P., Dolton, G., Clement, M., Llewellyn-Lacey, S., Price, D.A., 2012. A single autoimmune t cell receptor recognizes more than a million diff ;erent peptides. J. Biol. Chem. 287 (2), 1168–1177.

Xue, S., Gillmore, R., Downs, A., Tsallios, A., Stauss, H.J., 2005. Exploiting t cell receptor genes for cancer immunotherapy. Clin. Exp. Immunol. 139 (2), 167–172.

Yuval, E., Zachary, S., C. C. G, Thierry, M., W. A. M, 2018. Predicting the spectrum of tcr repertoire sharing with a data-driven model of recombination. Immunol. Rev. 284 (1), 167–179.

Zhou, C., Yu, H., Ding, Y., Guo, F., Gong, X.J., 2017. Multi-scale encoding of amino acid sequences for predicting protein interactions using gradient boosting decision tree. PLoS One 12 (8), e0181426.