

Received September 9, 2019, accepted October 13, 2019, date of publication October 18, 2019, date of current version October 30, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2948178

Prediction of Epitope-Associated TCR by Using Network Topological Similarity Based on Deepwalk

JINGSHU BI¹, YUANJIE ZHENG^{1,2,3,4}, FANG YAN¹, SUJUAN HOU¹, AND CHENGJIANG LI⁵

¹School of Information Science and Engineering, Shandong Normal University, Jinan 250358, China

²Key Lab of Intelligent Computing and Information Security in Universities of Shandong, Shandong Normal University, Jinan 250358, China

³Shandong Provincial Key Laboratory for Novel Distributed Computer Software Technology, Shandong Normal University, Jinan 250358, China

⁴Institute of Biomedical Sciences, Shandong Normal University, Jinan 250358, China

⁵Department of Electrical Engineering Information Technology, Shandong University of Science and Technology, Jinan 250031, China

Corresponding author: Yuanjie Zheng (zhengyuanjie@gmail.com)


This work was supported in part by the National Natural Science Foundation of China under Grant 61572300, Grant 81871508, and Grant 61773246, in part by the Taishan Scholar Program of Shandong Province of China under Grant TSHW201502038, and in part by the Major Program of Shandong Province Natural Science Foundation under Grant ZR2018ZB0419.

ABSTRACT Currently, there are many tools available online for T-cell epitope prediction. They usually focus on the binding of peptides to major histocompatibility complex (MHC) molecules on the surface of antigen-presenting cells (APCs). However, the binding of peptides and MHC complexes to the T-cell receptor (TCR) is also critical for the immune process. Identifying the binding of human epitopes to TCRs will be useful for developing vaccines. It also has great prospects in medical issues such as cancer and autoimmune diseases. We propose a similarity-based TCR-epitope prediction method using a similarity measure. This paper introduces the Deepwalk method to calculate the topological similarity between TCR-TCRs, constructs a TCR similarity network topology, and predicts the correlation between TCRs and epitopes based on known TCR-epitope associations. We selected data from 22 types of epitopes from the VDJDB database and trained models to implement TCR-epitope prediction. We trained a model on the data from the 22 types of epitopes, predicting which epitope each TCR belongs to. To compare with other methods, we also generated a second method involving training a model for each type of epitope so that we can predict which TCR is bound to the epitope from a large pool of TCRs. We used the ROC curve, PR curve and other evaluation indicators to evaluate our model in 10-fold cross-validation. In the first model, the AUC value of our method is 0.926, and that of the support vector machine (SVM) method is 0.924. Considering that no one has ever used the first prediction model, we used the second method for the predictions. The results show better predictive performance compared to the SVM method, TCRGP method and random forest method. Our AUC values range from 0.660 to 0.950. The experimental results show that our method outperforms other methods in TCR-epitope prediction, which can help predict the TCR-epitope.

INDEX TERMS Deepwalk, TCR-epitope associations, TCR-epitope prediction, similarity measure.

I. INTRODUCTION

The T-cell receptor (TCR) is a characteristic molecular marker found on the T-cell surface. Most TCRs consist of α and β chains [1], and a small number consist of δ and γ chains [2]. It is the gene recombination of the DNA fragment on both chains that produces a large number of different TCR sequences [3]–[5]. Epitopes are chemical groups

The associate editor coordinating the review of this manuscript and approving it for publication was Victor Hugo Albuquerque .

that have a specific structure on the surface of an antigen that determines antigen specificity. An antigen binds to a specific TCR through recognition of the epitope. The antigen is degraded into a polypeptide and specifically binds to the corresponding major histocompatibility complex (MHC) molecule. The MHC molecules then present antigenic peptides to the TCRs. If the MHC-peptide complex can be specifically recognized by the TCR, then it can induce the immune response of the T cell. Therefore, the binding of the TCR epitope is crucial in the immune process.

After TCR-epitope-specific recognition, T cells produce an effector T cell killing antigen cell with specific immune function. Medical research on the immune response mechanism of T cells has been useful for the development of targeted drugs and vaccines against infectious diseases [6]–[10]. An injected vaccine acts an antigenic substance, triggering an immune response with TCR-specific recognition and producing antibodies to eliminate the virus. The principle of identification between the TCR and the epitope is also an important part of vaccine development.

With the recent development of high-throughput sequencing technology [11], TCR-sequencing technology enables us to sequence a person's whole repertoire. We are able to select T cells' RNA or DNA or TCR-specific CDR3 regions for sequencing. Some tools can process sequences into a quantitative list, so we can use the sequencing results to perform epitope prediction-related experiments. In recent years, bioinformatics [12] has played a very beneficial role in the recognition of T-cell epitopes. Therefore, the scientific community is interested in further developing TCR sequencing and analysis tools. There are a great number of published TCR-peptide datasets available online, such as VDJDDB [13] and McPAS [14] containing information on different types of epitopes. These data make the analysis of TCR-peptide associations more comprehensive, making it possible for us to complete TCR-epitope prediction tasks.

Traditional T-cell epitope prediction often uses peptide scanning [15]. However, due to the enormous complexity of the immune process, this work is time consuming and costly. The use of computer-effective predictions of T-cell epitopes can avoid these problems, making accurate predictions that can then be proven by experiments. Therefore, the methods developed through bioinformatics have received extensive attention from researchers. Currently, in bioinformatics, research methods on epitope prediction are mainly divided into two groups: MHC-epitope prediction and TCR-epitope prediction. Most online tools are based on MHC-epitope predictions [16], mainly including methods based on molecular modeling [17], combined motif methods [18], quantitative matrixes [19] and machine learning [20]. Molecular modeling can reveal interaction mechanisms within molecules but is not suitable for high-throughput data processing. The binding motif method is simple and suitable for epitope prediction with less experimental data, but the complexity of MHC [21] binding to antigen peptides results in low prediction accuracy for the binding motif method. Methods involving quantitative matrixes are greatly improved compared to the combined motif method. The binding score of any epitope is obtained by using a specific scoring matrix (PSSM), but the effect of the overall peptide structure on the binding is neglected; furthermore, quantitative matrix-based methods also suffer from data-overfitting issues. Machine-learning methods include Artificial Neural Network (ANN) [22] and support vector machine (SVM) [23]. The SVM method is widely used in MHC-epitope identification and other fields.

It is a traditional machine learning that results in better prediction results for MHC-epitope identification than other machine-learning methods. However, in the identification of TCR epitopes, the SVM classification prediction method has never been used. Therefore, in this article, we use the traditional method of SVM for classification prediction of TCR epitopes. There are many web tools that can predict combinations of MHC epitopes, and existing predicting tools have achieved high accuracy. However, there are few tools for predicting the combination of TCR-epitope. Currently used TCR-epitope prediction methods consist of TCRGP [24], TCRdist [25] and a random forest-based prediction method [26]. TCRGP is a Gaussian process classification based on non-parametric models. TCRdist is a method for calculating distances to determine whether a TCR is closer to an epitope-specific cluster or background TCR cluster. The random forest method predicts the identification of the TCR epitope by extracting the features of TCR amino acid sequences and placing them into a random forest classifier.

However, none of these prediction methods takes the topological structure of a TCR-epitope network into account. The TCR-epitope topology network based on the TCR-epitope data verified by known biological experiments and the calculated TCR sequence similarity contributes to the TCR-epitope prediction. Many studies have proposed methods for calculating the similarity of TCR sequences: TCRdist distance measure [25]: Using the BLOSUM62 scoring matrix to calculate the similarity weight mismatch distance between two receptor loops; GLIPH Algorithm [27]: Searches for CDR3 global similarity and local similarity in the CDR region of TCR to cluster receptors. CDRdist [28]: Calculates the similarity score of TCR sequences based on sequence alignment and finally standardizes the scores. Profile distance measure [26]: Constructs profiles based on physicochemical properties such as hydrophobicity, alkalinity, and helicity of each amino acid; then, calculated the distance between two profiles by using a weighted Euclidean distance. By calculating the similarity of the TCR sequences, we found that the TCR distance similarity scores belonging to the same type of epitope were lower than the heterogeneous scores. Based on this finding, we constructed a TCR similarity topology by the TCR similarity score in TCR-epitope prediction. The experimental results prove that the TCR similarity topology contributes to the TCR-epitope prediction.

In this paper, we proposed a method based on TCR-TCR topology similarity. Constructing a TCR-TCR similarity topology space by calculating the distance between TCR sequences. Then, using a deep learning method, Deepwalk [29], we extracted the features of the vertices in the topology structure, and then predicted the TCR epitope based on the known data. We used the ROC curve, PR curve and other evaluation indicators to evaluate our prediction model under 10-fold cross-validation. The experimental results show that our method makes contributions to TCR-epitope prediction. Figure 1 shows the core part of our method.

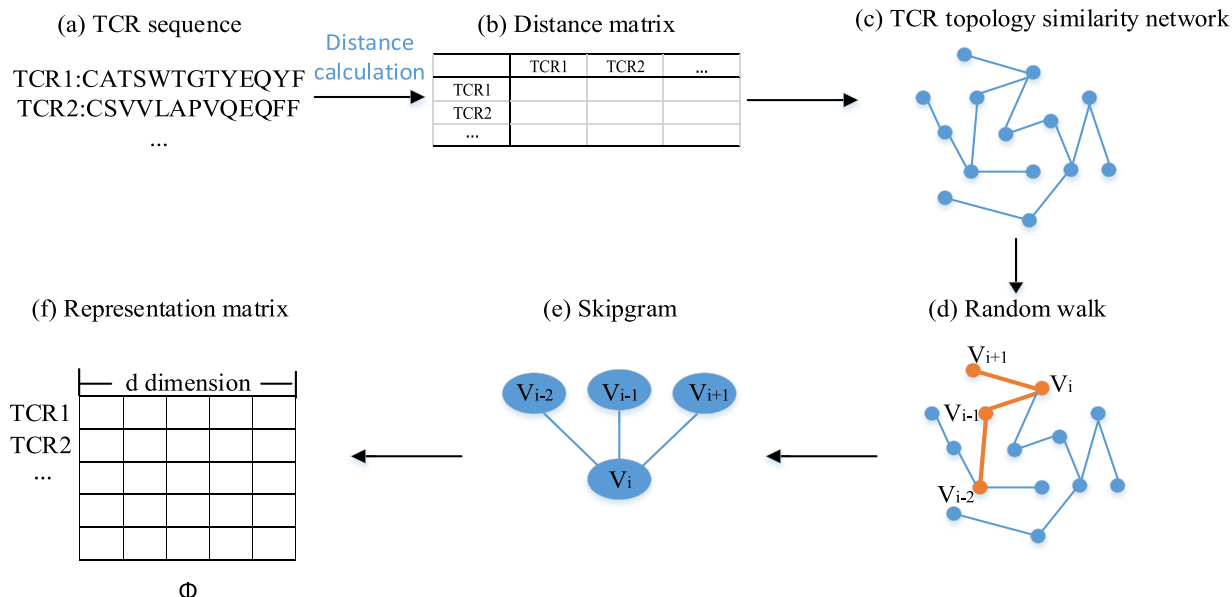


FIGURE 1. The core part of TCR-epitope predictions in our method.

II. MATERIALS

A. DATA

Most TCRs are composed of α and β chains; the rest are composed of δ and γ chains. At present, there are many available data for the α chain and β chain online. Our dataset gathers 3654 TCR sequences with 22 types of epitopes from the VDJDDB (<https://vdjdb.cdr3.net>), which compiles information from many different studies and contains TCR sequences from many different individuals. TCR β sequence data contains V-region, J-region and CDR3 amino acid sequence information. Each sequence in the VDJDDB gives a confidence score (0-3). Due to the large difference in the number of TCRs of different types of epitopes in the VDJDDB database, a maximum of 413 TCRs, but the least is only one. To balance the difference in sample size, all epitopes we selected contain at least 50 TCR sequences [24] with a VDJ score greater than 1. As is known to us all that the TCR CDR3 β sequence plays a significance role in the process of recognizing peptides. Therefore, in our experiment, we only selected the CDR3 β sequence and deleted the data from the VDJDDB that did not meet our requirements. A detailed description of the epitope dataset is presented in Table 1.

In the final comparison experiment, we selected some background TCRs from Dash *et al.* [25]. Moreover, we ensured that the number of epitope-specific TCRs and background TCRs were the same size [24] in the training sets and test sets to ensure the authenticity of the model evaluation and to observe how the trained model adapts to the new data.

B. SEQUENCE REPRESENTATION

TCR sequences are composed of 20 amino acids, each of which needs to be characterized so that it can then be used by computational methods. Kidera *et al.* [30]

TABLE 1. Dataset collected from the VDJDDB database, which contains 22 types of epitopes for 8 viruses.

Epitope gene	Epitope	Number of sequences
BMLF1	GLCTLVAML	299
BZLF1	RAKFKQLL	225
BRLF1	YVLDHLIVV	66
M1	GILGFVFTL	239
HA	PKYYKQNTLKLAT	56
VP22	RPRGEVRFLL	68
NS3	ATDALMTGY	152
NS3	CINGVCWTV	76
NS3	KLVALGINAV	65
NS4B	LLWNGPMAV	223
P24	EIYKRWII	81
P24	FRDYVDRFYKTLRAEQASQE	141
P24	GPGHKARVL	62
P24	KAFSPEVIPMF	234
P24	KRWIILGLNK	212
P24	TPQDLNNTML	52
Nef	FLKEKGGGL	104
p65	IPSNVHHY	65
p65	NLVPVAVTV	413
p65	TPRVGGGAM	184
NS3	GTSGPSIVNR	65
NS3	GTSGPSIINR	51

extracted 10 orthogonal factors from 188 physical properties of 20 amino acids by multivariate statistical analysis. These 10 factors with appropriate weighting factors can represent most of the 188 physical properties. These factors include helicity, hydrophobicity, and structural preferences, etc. In our experiment, each amino acid in the TCR sequence was encoded by these 10 orthogonal factors. We encoded the physicochemical properties of amino acids and extracted the features of each TCR sequence. Because each TCR sequence has a different length, the variables for each amino acid input are not equal.

III. METHODS

A. TCR-TCR SIMILARITY SPACE

In this section, we use a distance-based method such as physicochemical differences or sequence alignment-gap penalty for each amino acid to construct a topological

TABLE 2. Examples of the GapAlign score for the TCR sequences of the ATDALMTGY epitope.

ATDALMTGY \ ATDALMTGY	CAISESTVGNQPQHF	CAISESTVGNQPQHF	CAISESATGYQPQHF	CAISESSAGNQPQHF	CAISESSIGNQPQHF
CAISEGAMGNQPQHF	33	33	36	33	30
CAISEEGIGNQPQHF	27	27	48	36	24
CAAADDEEIGNQPQHF	42	42	63	51	39
CAIGDDRAGNQPQHF	54	54	66	42	54
CAISDQTSIGNQPQHF	30	30	51	36	39

TABLE 3. Examples of the GapAlign score for the TCR sequences of the ATDALMTGY and GILGFVFTL epitopes.

ATDALMTGY \ GILGFVFTL	CASSRSRAGELFF	CASSRSRANEQFF	CASSFRSTDTQYF	CASSKRSTDTQYF	CASSRSSYEQYF
CAISEGAMGNQPQHF	100	100	109	112	97
CAISEEGIGNQPQHF	88	88	97	100	85
CAAADDEEIGNQPQHF	88	88	97	100	85
CAIGDDRAGNQPQHF	85	85	97	100	85
CAISDQTSIGNQPQHF	85	85	94	97	82

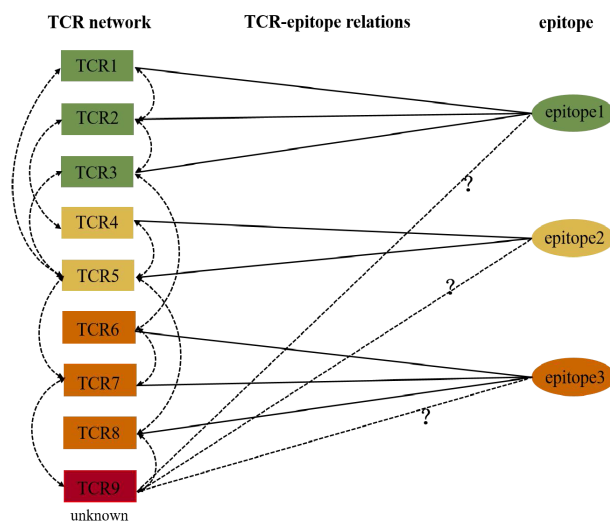


FIGURE 2. TCR-epitope bipartite network.

similarity space for TCR sequences. The TCR-epitope bipartite network is shown in Figure 2. We construct a TCR sequence-topology similarity space. TCRs of the same color indicate that they are associated with the same type of epitope. The dashed line indicates that a distance-based approach is used to construct a topology similarity network for TCR sequences. The solid line indicates the known relationship between TCR and epitope. Our goal is to predict which epitope is associated with the unknown TCR.

Inspired by Meysman *et al.* [31], in this article, we mainly used the GapAlign method to calculate the distances between TCR sequences. The lower the GapAlign score, the higher the similarity between the two TCR sequences. From the GapAlign scores in Table 2 and Table 3, it can be seen that the similarity of TCR sequences in the same type of epitope is higher than the similarity of TCR sequences belonging to different epitopes. Therefore, we constructed a TCR-similarity topological network based on the differences in sequence similarity within and between epitopes to help predict the relationship between the TCR and the epitope. In section IV, we added the Profile and CDRdist distance calculation methods to compare with the GapAlign method. The experimental

results show that the GapAlign method outperformed than other two methods.

GapAlign: This distance originated from the TCRdist method proposed by Dash *et al.*, which is based on sequence alignment and uses the BLOSUM90 [32] scoring matrix.

Profile: This distance is used to calculate the physicochemical properties between two TCR sequences. We normalized the alkalinity, hydrophobicity, and helicity values of each amino acid, constructed a profile of the TCR CDR3 region, and calculated the distance of the two profiles using weighted Euclidean distances. The sum of the three profile distances is the final score.

CDRdist: This distance performs local alignment using the Smith-Waterman algorithm [33]. It normalizes to the [0,1] interval by using the algorithm for local alignment and dividing the alignment score by the minimum of the two sequence self scores.

B. DEEPWALK SIMILARITY LEARNING

Deepwalk [29] is a deep learning method that vectorizes vertices in the graph and represents the potential relationship of the vertices. This method obtains local information about the random walk by maximizing the co-occurrence probability of the vertex v_j of the visited node in the window w to learn the vector representation of the vertex. Deepwalk contains two main steps: (1) random walk, (2) an update step with a skipgram algorithm [34]. The random walk generator takes a random vertex v_j in graph G as the root of random walk w_{v_j} . In our experiment, the walk length t of the random walk is fixed. A walk sequence randomly selects the neighbor of the last visited node until it reaches the maximum length t . For each random walk sequence, it maximizes the co-occurrence possibility of vertices in a window size w , calculated as follows:

$$Pr(\{v_{j-w}, \dots, v_{j+w}\} \setminus v_j | \phi(v_j)) = \prod_{i=j-w, i \neq j}^{j+w} Pr(v_i | \phi(v_j)) \quad (1)$$

The skipgram algorithm iterates all possible matches of the random walk sequence in window w . For each j , $\phi(v_j)$ means vertex v_j maps to its representation space; $\phi \in R^{|\mathcal{V}| \times d}$ is represented by a matrix, where d is the embedding size,

and $|v|$ is the set of vertices. After giving v_j a representation in space, we want to maximize the probability of its neighbors in the walk sequence. We can use a variety of models to learn the posterior distribution, but these models require huge computing resources. The calculation of $Pr(v_i|\phi(v_j))$ is not feasible. Therefore, hierarchical softmax [35] is introduced to factorize $Pr(v_i|\phi(v_j))$. We assign the vertices in the walk sequence to the leaves of the binary tree and transform the prediction problem into a hierarchy to maximize the specific path using vertices as leaf nodes of the Huffman tree to shorten training time. The formula of $Pr(v_i|\phi(v_j))$ is as follows:

$$Pr(v_i|\phi(v_j)) = \prod_{k=1}^{\lceil \log|V| \rceil} 1/(1 + e^{-\phi(v_j) \cdot \varphi(b_k)}) \quad (2)$$

where $\varphi(b_k) \in R^d$ represents the parent node of tree node b_k . The path to v_i is identified by the sequence of tree nodes $(b_0, b_1, \dots, b_{\log|v|})$, where $b_0 = \text{root}$ and $b_{\log|v|} = v_i$. The Huffman tree we introduced can speed up training time by assigning shorter paths to frequent vertices in a random walk.

We use distance methods to construct the TCR-TCR similarity topology and input the obtained TCR sequence similarity relation matrix into the deepwalk method. Starting from a random vertex v_j (the TCR sequence), a sequence of local relationships between TCR-TCRs is obtained by a random walk. The walk length t , walk per vertex γ , window size w and embedding size d are appropriately adjusted according to the data. Then, the skipgram algorithm iteratively updates the representation of the vertices by maximizing the co-occurrence probability and finally obtains the vector matrix ϕ that is learned. The vector matrix ϕ is the potential representation of the relationship between the vertices in the topology similarity network learned by deepwalk. Each row is equivalent to the feature of a TCR sequence. We combine the features extracted from the TCR topology with the original feature representation of the sequence described in Section II then place them into the SVM classifier to predict the TCR-epitope relationships.

C. CLASSIFICATION ALGORITHM AND VALIDATION

In this paper, we use the libsvm package [36] to perform TCR-epitope prediction. The SVM Classifier maps the input vector to a high-dimensional space and uses an optimal hyperplane to maximize the separation of a given set of training data. We use the RBF kernel, constantly changing the value of the error penalty (C) and gamma (g) ($C=10, g=0.012$) by grid search to maximize the prediction accuracy. We combined the TCR sequence features obtained by topological similarity learning with the amino acid features encoded, each type of epitope is a label. The goal of SVM is to construct a classifier to classify TCR sequences for 22 types of epitopes.

In the next section, we compare our methods with TCRGP and random forest methods. TCRGP is a Gaussian process

classification algorithm that predicts whether TCR-epitope can specifically bind. It based on sequence similarity and modeling sequences by kernel function. The random forest method is a common classification method that can predicts the identification of the TCR epitope.

We use 10-fold cross-validation to train the SVM model. In our experiment, the TCR sequence with the epitope label was randomly divided into 10 disjoint subsets; one subset was used for testing, and the other 9 subsets were subjected to multiple iterations training. This step was repeated 5 times and we selected the average of 5 results as the final result. To evaluate the performance of our method, the ROC curve and PR curve, as well as accuracy, sensitivity, and precision, were calculated to evaluate the quality of the model:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (4)$$

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

where TN, TP, FN and FP represent true negative, true positive, false negative, and false positive values, respectively. The ROC curve is drawn based on these indicators. The area under the ROC curve (AUC) can be used as a reliable indicator of the evaluation of the model.

IV. EXPERIMENT AND RESULT

A. TCR TOPOLOGY PREDICTION

We added the features of the TCR similarity topology structure based on the SVM method. By calculating the GapAlign score of the TCR sequences, we found that there is a certain degree of correlation between the TCR sequences of the same type of epitope and considered that there is a similar relationship between TCR sequences. Therefore, we used the GapAlign distance calculation method (threshold=90) for constructing the TCR topology similarity network and use the deepwalk method for similarity learning.

Deepwalk has a total of 4 parameters. We selected a set of optimal parameters through grid search algorithm. The best performance of deepwalk is achieved at window size $w=5$, walk length $t=40$, dimension $d=32$ and number of walks $\gamma = 10$. Therefore, we chose this set of values as the default parameters for deepwalk. We compared our method with the SVM method in Figure 3 and Table 4. Deepwalk extracts similarity topology features between TCR sequences. Figure 3 clearly shows that the value of the AUC for our method is higher than that of the SVM method. In Table 4, we used the three evaluation indicators mentioned above to directly reflect the performance of the two models. Compared with the SVM method, our method has improved performance, achieving a precision of 0.720, a sensitivity of 0.608, and an accuracy of 0.706. The results show that adding the TCR similarity topology feature can help predict the epitope.

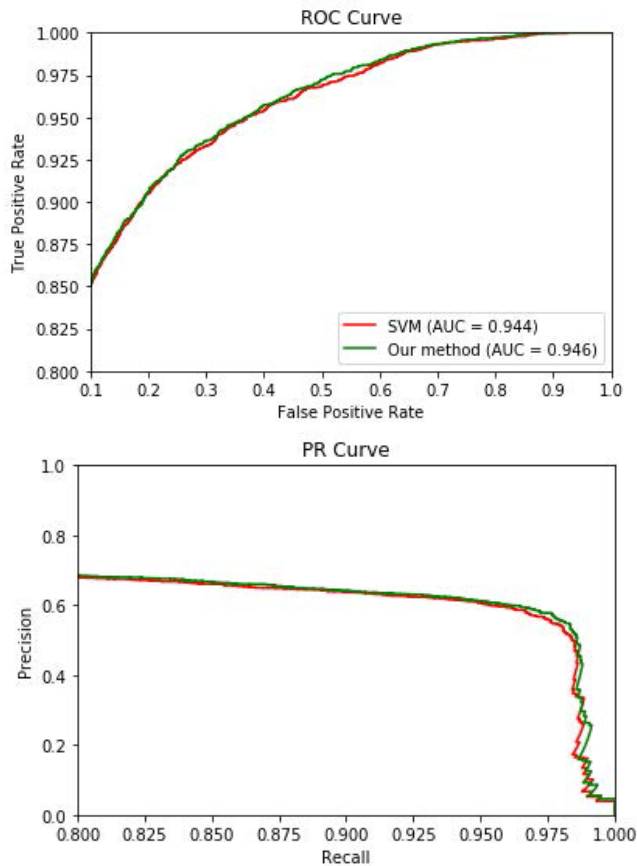


FIGURE 3. ROC curves and PR curves for the two methods. SVM method (red line) vs Our method (green line).

TABLE 4. Results from the prediction methods.

	Precision	Sensitivity	Accuracy
SVM	0.711	0.603	0.701
Our method	0.720	0.608	0.706

B. PREDICTION ACCURACY OF EPITOPES (TOP30, TOP40, TOP50)

In addition, for comparison, we ranked the posterior probabilities of each type of epitope with the two methods. The higher the posterior probability, the greater the association between the TCR sequence and that epitope. We sorted the posterior probabilities of each type of epitope prediction, selected the top 30, top 40, and top 50 TCRs, and then average the prediction accuracy of the 22 types of epitopes. Figure 4 summarizes our method and the SVM method for predicting TCR epitope capabilities. Specifically, the average predicted values of the 22 types of epitopes are 0.926, 0.893, and 0.856 for the top-ranked 30, 40 and 50 TCRs in our method, respectively. In contrast, these respective values are 0.925, 0.890, and 0.855 when ranking TCRs by the SVM method. Larger proportions of top-ranked TCRs indicate that our model is better than the SVM method. These results indicate that constructing a TCR similarity topology network contributes to TCR epitope prediction.

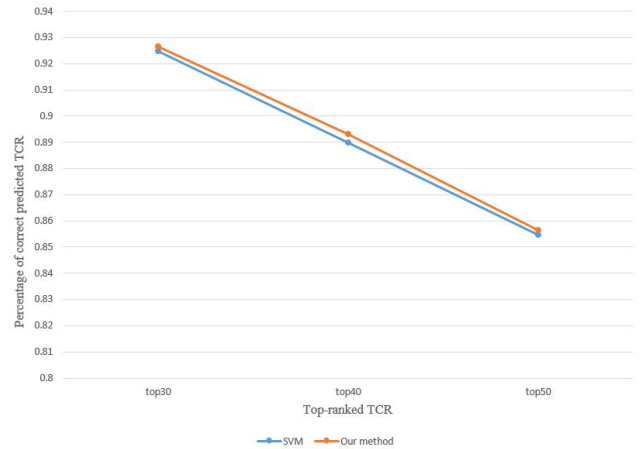


FIGURE 4. Ranking TCR based on the posterior probability of each type of epitope. We select the top 30, top 40 and top 50 TCRs for each epitope separately and obtain the average prediction accuracy.

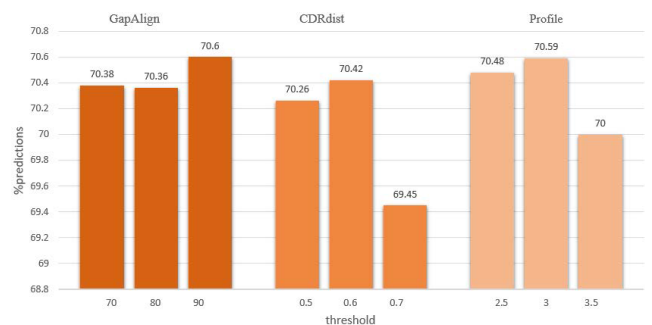


FIGURE 5. Comparison of different thresholds for different distance and distance measures.

C. DIFFERENT DISTANCE MEASURES AND THRESHOLDS

Experimenting with different thresholds for the three different distance metrics mentioned above, we can explore which distance measure or which measure threshold has better performance.

In Figure 5, we can see the overall prediction of the three thresholds; the GapAlign method is superior to the other two methods. This is also proven in the article by Meysman et al. [31]. The principles of the three distance methods are different. The CDRdist method is similar to the GapAlign method; both are based on sequence alignment and incorporate the BLOSUM replacement matrix. The difference is that the GapAlign method deeply analyzes the internal structure of the TCR α and TCR β chains. It then calculates the alignment of CDR2.5 α , CDR3 α , CDR2.5 β , etc., and sets a reasonable weight for each region’s sequence. CDRdist only considers the sequence alignment of the CDR3 region. The Profile method is based on the physicochemical properties of the amino acids, such as hydrophobicity, alkalinity, and helicity, makes these values Z-normalized to construct a profile, and calculates the distance between two profiles. From the experimental results, the GapAlign method has a good predictive effect.

TABLE 5. Prediction results of Our method, SVM, TCRGP and Random Forest with 10-fold cross-validation.

Epitope	No.of associated TCRs	AUC			
		Deepwalk	SVM	TCRGP	RF
IPSNVHHY	65	0.848	0.824	0.864	0.839
NLVPMVATV	413	0.904	0.901	0.893	0.910
TPRVTGGGAM	184	0.934	0.937	0.936	0.964
GLCTLVAML	299	0.933	0.927	0.927	0.926
RAKFKQLL	225	0.886	0.878	0.892	0.879
YVLDHLIVV	66	0.723	0.699	0.677	0.730
GILGFVFTL	239	0.900	0.902	0.887	0.915
PKYVKQNTLKLAT	56	0.739	0.723	0.669	0.601
ATDALMTGY	152	0.883	0.882	0.880	0.930
CINGVCWTV	76	0.890	0.889	0.880	0.878
KLVALGINAV	65	0.694	0.683	0.685	0.665
RPRGEVRF	68	0.933	0.926	0.928	0.902
LLWNGPMAV	223	0.836	0.828	0.830	0.825
GTSGPSVNR	65	0.892	0.889	0.794	0.856
GTSGSPINR	51	0.843	0.832	0.810	0.742
EYKRWII	81	0.879	0.879	0.835	0.842
FRDYVDRFYKTLRAEQASQE	141	0.950	0.946	0.940	0.993
GPGHKARVL	62	0.660	0.636	0.744	0.594
KAFSPEVIPMF	134	0.865	0.873	0.856	0.852
KRWIILGLNK	212	0.875	0.872	0.913	0.906
TPQDLNTML	52	0.818	0.809	0.856	0.898
FLKEKGGL	104	0.882	0.883	0.886	0.870
Average AUC		0.853	0.846	0.844	0.842

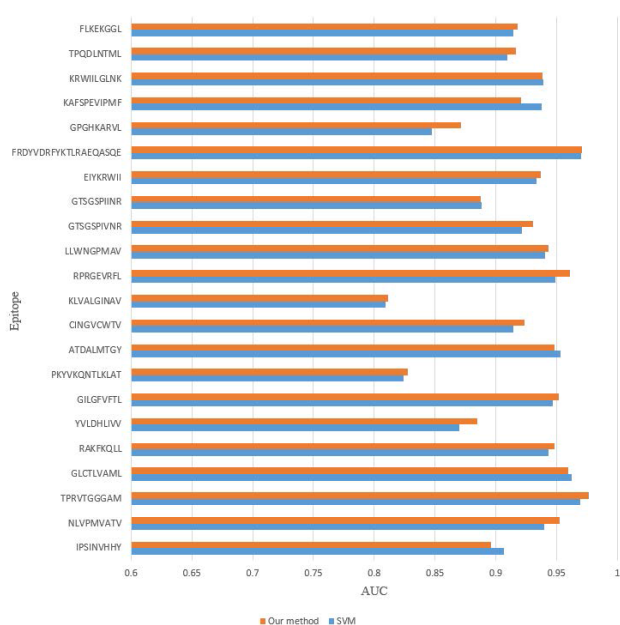


FIGURE 6. AUC value for each type of epitope.

D. COMPARISON OF EACH TYPE OF EPITOPE

We also compared the AUC values of each type of epitope for the two prediction methods. Figure 6 shows that the AUC value of 16 of the 22 types of epitopes is higher than that of the SVM method. Poor prediction results for some of the types of epitopes is thought to be caused by one or more of the following reasons: (1) The TCR sequences in each type of epitope are different, and there are also significant differences between sequences. This similarity is related to structure and function and is also closely related to the amino acid of each locus. The amino acid of each locus affects the nature of the amino acids around it. (2) Our data have

a sample imbalance problem that affects the performance of our models and the accuracy of our predictions. We are considering whether we can reduce the impact of this imbalance. (3) Because we considered the topology of the TCR sequences, the construction of the topology involves the distance method and the selection of their thresholds. Different distance metrics and their choice of thresholds will affect the experimental results. In this comparison experiment, we still use the GapAlign method (threshold=90) to construct the TCR-TCR similarity network topology.

E. COMPARISONS TO OTHER METHODS

Considering that no one has performed the prediction model described above, in order to compare the performance of our model, we trained a model for each type of epitope. We randomly selected the background TCRs, with an equal number of epitope-specific and background TCRs for each type of epitope.

Because there are few TCR-epitope prediction methods, we compared our method with only two methods, the SVM, TCRGP and random forest method, to each type of epitope prediction (Table 5). Unfortunately, our background data from [16] did not contain information about the J gene that the random forest classifier needs. However, according to the description of [18], the weights they ascribe to the J genes are not very high. Therefore, in running the random forest classifier, we only used the V gene. The experimental data set consists of the CDR3β sequence data, the dataset we described in II. All four methods use the same sequence information and 10-fold cross-validation. Table 5 shows that the AUC value of 9 of the 22 types of epitopes was higher than that of the other methods. The average AUC value achieved by our method is 0.853, with a minimum of 0.660 for epitope GPGHKARVL and a maximum of 0.950 for FRDYVDRFYKTLRAEQASQE.

V. CONCLUSION

Identifying the TCR-epitope associations is important for exploring the immune response mechanism of the human body and further improving the development of medical vaccines. In this paper, we propose a similarity-based TCR-epitope prediction method using a topological similarity structure. First, we construct a TCR similarity topological network by calculating the distance between TCR sequences. Second, we use the deepwalk method to extract the features of the TCR similarity topology network and then add it to the original amino acid features. Finally, we put these features into an SVM classifier for classification prediction, where ranking the posterior probabilities of each type of epitope gives us the final result. Based on 10-fold cross-validation, we used ROC curve, PR curve and other evaluation indicators to evaluate our prediction model.

Although we have achieved good results by adding the TCR similarity topological network to the TCR-epitope prediction, there are also some inevitable limitations. First, to improve the prediction accuracy of our method, the CDR α chain or TCR structure similarity can be added for prediction so that the features of the TCR are more complete and may improve the prediction accuracy. The currently known TCR-epitope associations are insufficient. Therefore, TCR-gene and gene-epitope associations can be used for similarity learning, which may improve the predicted results. Next, in this article, we only consider the similarity topology of TCRs; we can also construct the similarity topology of epitopes to improve our prediction.

REFERENCES

- [1] P. Dash, J. L. McClaren, T. H. Oguin, III, W. Rothwell, B. Todd, M. Y. Morris, J. Becksfors, C. Reynolds, S. A. Brown, P. C. Doherty, and P. G. Thomas, "Paired analysis of TCR α and TCR β chains at the single-cell level in mice," *J. Clin. Invest.*, vol. 121, no. 1, pp. 288–295, 2011.
- [2] T. J. Allison, C. C. Winter, J.-J. Fournié, M. Bonneville, and D. N. Garboczi, "Structure of a human $\gamma\delta$ T-cell antigen receptor," *Nature*, vol. 411, no. 6839, pp. 820–824, 2001.
- [3] C. H. Bassing, W. Swat, and F. W. Alt, "The mechanism and regulation of chromosomal V(D)J recombination," *Cell*, vol. 109, no. 2, pp. S45–S55, 2002.
- [4] T. P. Arstila, A. Casrouge, V. Baron, J. Even, J. Kanellopoulos, and P. Kourilsky, "A direct estimate of the human $\alpha\beta$ T cell receptor diversity," *Science*, vol. 286, no. 5441, pp. 958–961, 1999.
- [5] H. S. Robins, P. V. Campregher, S. K. Srivastava, A. Wachter, C. J. Turtle, O. Khsai, S. R. Riddell, E. H. Warren, and C. S. Carlson, "Comprehensive assessment of T-cell receptor β -chain diversity in $\alpha\beta$ T cells," *Blood*, vol. 114, no. 19, pp. 4099–4107, 2009.
- [6] W. Fischer, S. Perkins, J. Theiler, T. Bhattacharya, K. Yusim, R. Funkhouser, C. Kuiken, B. Haynes, N. L. Letvin, B. D. Walker, B. H. Hahn, and B. T. Korber, "Polyvalent vaccines for optimal coverage of potential T-cell epitopes in global HIV-1 variants," *Nature Med.*, vol. 13, no. 1, pp. 100–106, 2007.
- [7] S. L. Elliott, A. Suhrbier, J. J. Miles, G. Lawrence, S. J. Pye, T. T. Le, A. Rosenstengel, T. Nguyen, A. Allworth, S. R. Burrows, J. Cox, D. Pye, D. J. Moss, and M. Bharadwaj, "Phase I trial of a CD8 $^{+}$ T-cell peptide epitope-based vaccine for infectious mononucleosis," *J. Virol.*, vol. 82, no. 3, pp. 1448–1457, 2008.
- [8] A. Patronov and I. Doytchinova, "T-cell epitope vaccine design by immunoinformatics," *Open Biol.*, vol. 3, no. 1, 2013, Art. no. 120139.
- [9] D. H. Barouch, K. L. O'Brien, N. L. Simmons, S. L. King, P. Abbink, L. F. Maxfield, Y.-H. Sun, A. La Porte, A. M. Riggs, D. M. Lynch, S. L. Clark, K. Backus, J. R. Perry, M. S. Seaman, A. Carville, K. G. Mansfield, J. J. Szinger, W. Fischer, M. Muldoon, and B. Korber, "Mosaic HIV-1 vaccines expand the breadth and depth of cellular immune responses in rhesus monkeys," *Nature Med.*, vol. 16, no. 3, pp. 319–323, 2010.
- [10] Y. Tian, R. S. Antunes, and J. Sidney, "A review on T cell epitopes identified using prediction and cell-mediated immune models for *Mycobacterium tuberculosis* and *Bordetella pertussis*," *Frontiers Immunol.*, vol. 9, p. 2778, 2018.
- [11] S. Friedensohn, T. A. Khan, and S. T. Reddy, "Advanced methodologies in high-throughput sequencing of immune repertoires," *Trends Biotechnol.*, vol. 35, no. 3, pp. 203–214, 2017.
- [12] A. Lesk, *Introduction to Bioinformatics*. Oxford, U.K.: Oxford Univ. Press, 2019.
- [13] M. Shugay, D. V. Bagaev, I. V. Zvyagin, R. M. Vroomans, J. C. Crawford, G. Dolton, E. A. Komech, A. L. Sycheva, A. E. Koneva, E. S. Egorov, and A. V. Eliseev, "VDJdb: A curated database of T-cell receptor sequences with known antigen specificity," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D419–D427, 2017.
- [14] N. Tickotsky, T. Sagiv, J. Prilusky, E. Shifrut, and N. Friedman, "McPAS-TCR: A manually curated catalogue of pathology-associated T cell receptor sequences," *Bioinformatics*, vol. 33, no. 18, pp. 2924–2929, 2017.
- [15] J. Sidney, E. Assarsson, C. Moore, S. Ngo, C. Pinilla, A. Sette, and B. Peters, "Quantitative peptide binding motifs for 19 human and mouse MHC class I molecules derived using positional scanning combinatorial peptide libraries," *Immunome Res.*, vol. 4, no. 1, 2008, Art. no. 2.
- [16] M. G. Rudolph, R. L. Stanfield, and I. A. Wilson, "How TCRs bind MHCs, peptides, and coreceptors," *Annu. Rev. Immunol.*, vol. 24, pp. 419–466, Apr. 2006.
- [17] R. E. Soria-Guerra, R. Nieto-Gomez, D. O. Govea-Alonso, and S. Rosales-Mendoza, "An overview of bioinformatics tools for epitope prediction: Implications on vaccine development," *J. Biomed. Inform.*, vol. 53, pp. 405–414, Feb. 2015.
- [18] K. Falk, O. Rötzschke, S. Stevanović, G. Jung, and H.-G. Rammensee, "Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules," *Nature*, vol. 351, no. 6324, pp. 290–296, 1991.
- [19] M. Bhasin and G. P. S. Raghava, "A hybrid approach for predicting promiscuous MHC class I restricted T cell epitopes," *J. Biosci.*, vol. 32, no. 1, pp. 31–42, 2007.
- [20] G. L. Zhang, H. R. Ansari, P. Bradley, G. C. Cawley, T. Hertz, X. Hu, N. Jovic, Y. Kim, O. Kohlbacher, and O. Lund, "Machine learning competition in immunology-prediction of HLA class I molecules," *J. Immunol. Methods*, vol. 374, nos. 1–2, pp. 1–4, 2011.
- [21] R. M. Zinkernagel and P. C. Doherty, "The discovery of MHC restriction," *Immunol. Today*, vol. 18, no. 1, pp. 14–17, 1997.
- [22] C. Lundegaard, O. Lund, and M. Nielsen, "Prediction of epitopes using neural network based methods," *J. Immunol. Methods*, vol. 374, nos. 1–2, pp. 26–34, 2011.
- [23] M. Bhasin and G. P. S. Raghava, "Prediction of CTL epitopes using QM, SVM and ANN techniques," *Vaccine*, vol. 22, nos. 23–24, pp. 3195–3204, 2004.
- [24] E. Jokinen, M. Heinonen, J. Huuhtanen, S. Mustjoki, and H. Lähdesmäki, "TCRGP: Determining epitope specificity of T cell receptors," *bioRxiv*, Jan. 2019, Art. no. 542332.
- [25] P. Dash, A. J. Fiore-Gartland, T. Hertz, G. C. Wang, S. Sharma, A. Souquette, J. C. Crawford, E. B. Clemens, T. H. O. Nguyen, K. Kedzierska, N. L. La Gruta, P. Bradley, and P. G. Thomas, "Quantifiable features define epitope-specific T cell receptor repertoires," *Nature*, vol. 547, no. 7661, pp. 89–93, 2017.
- [26] N. De Neuter, W. Bittremieux, C. Beirnaert, B. Cuypers, A. Mrzic, P. Moris, A. Suls, V. Van Tendeloo, B. Ogunjimi, K. Laukens, and P. Meysman, "On the feasibility of mining CD8 $^{+}$ T cell receptor patterns underlying immunogenic peptide recognition," *Immunogenetics*, vol. 70, no. 3, pp. 159–168, 2018.
- [27] J. Glanville, H. Huang, A. Nau, O. Hatton, L. E. Wagar, F. Rubelt, X. Ji, A. Han, S. M. Krams, C. Pettus, N. Haas, C. S. L. Arlehamn, A. Sette, S. D. Boyd, T. J. Scriba, O. M. Martinez, and M. M. Davis, "Identifying specificity groups in the T cell receptor repertoire," *Nature*, vol. 547, no. 7661, pp. 94–98, 2017.

[28] N. Thakkar and C. Bailey-Kellogg, "Balancing sensitivity and specificity in distinguishing TCR groups by CDR sequence similarity," *bioRxiv*, Jan. 2019, 526467.

[29] B. Perozzi, R. Al-Rfou, and S. Skiena, "DeepWalk: Online learning of social representations," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2014, pp. 701–710.

[30] A. Kidera, Y. Konishi, M. Oka, T. Ooi, and H. A. Scheraga, "Statistical analysis of the physical properties of the 20 naturally occurring amino acids," *J. Protein Chem.*, vol. 4, no. 1, pp. 23–55, 1985.

[31] P. Meysman, N. De Neuter, S. Gielis, D. B. Thi, B. Ogunjimi, and K. Laukens, "The workings and failings of clustering T-cell receptor beta-chain sequences without a known epitope preference," *bioRxiv*, Jan. 2018, Art. no. 318360.

[32] S. Henikoff and J. G. Henikoff, "Amino acid substitution matrices from protein blocks," *Proc. Nat. Acad. Sci. USA*, vol. 89, no. 22, pp. 10915–10919, 1992.

[33] M. S. Waterman, T. F. Smith, and W. A. Beyer, "Some biological sequence metrics," *Adv. Math.*, vol. 20, no. 3, pp. 367–387, 1976.

[34] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*. [Online]. Available: <https://arxiv.org/abs/1301.3781>

[35] A. Mnih and G. E. Hinton, "A scalable hierarchical distributed language model," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1081–1088.

[36] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, 2011, Art. no. 27.



FANG YAN received the M.Sc. degree in computer science from Shandong Normal University, China, in 2016, where she is currently pursuing the Ph.D. degree. Her research interests include information biology, image processing, and machine vision.



SUJUAN HOU received the Ph.D. degree from Chongqing University, Chongqing, China, in 2015. She is currently an Associate Professor with the School of Information Science and Engineering, Shandong Normal University, Jinan, China. Prior to that, she was a Visitor with the Faculty of Engineering and Information Technology (FEIT), University of Technology Sydney (UTS), from 2013 to 2015. Her research interests include video representation, and multimedia analysis and processing.



JINGSHU BI was born in Shandong, China, in 1996. She is currently pursuing the master's degree with the School of Information Science and Technology, Shandong Normal University. Her research interests include computational biology and bioinformatics.



YUANJIE ZHENG was a Senior Research Investigator with the Perelman School of Medicine, University of Pennsylvania. He is currently a Professor with the School of Information Science and Engineering, Shandong Normal University, and a Taishan Scholar of People's Government, Shandong, China. He is also serving as the Vice Dean of the School of Information Science and Technology and the Institute of Life Sciences, Shandong Normal University. His research interests

include medical image analysis, translational medicine, computer vision, and computational photography, enhance patient care by creating algorithms for automatically quantifying and generalizing the information latent in various medical images for tasks, such as disease analysis and surgical planning through the applications of computer vision and machine learning approaches to medical image analysis tasks and the development of strategies for image-guided intervention/surgery.



CHENGJIANG LI is currently an Instructor with the Shandong University of Science and Technology. His research interests include image analysis, computer vision, and artificial intelligence.

...