INeo-Epp: A novel T-cell HLA class-I immunogenicity or neoantigenic epitope prediction method based on sequence related amino acid features

Guangzhi Wang[1,2], Huihui Wan[2,3], Xingxing Jian[2,4], Yuyu Li[1], Jian Ouyang[2], XiaoxiuTan[3], Yong Zhao[1*], Yong Lin[3*], Lu Xie[1,2*]

[1] College of Food Science and Technology, Shanghai Ocean University, Shanghai, 201306, China
[2] Shanghai Center for Bioinformation Technology, Shanghai Academy of Science and Technology, Shanghai, 201203, China
[3] School of Medical Instrument and Food Engineering, University of Shanghai for Science and Technology, Shanghai, 200093, China
[4] Key Laboratory of Carcinogenesis and Cancer Invasion, Ministry of Education; Key Laboratory of Carcinogenesis, National Health and Family Planning Commission, Xiangya Hospital, Central South University, Changsha,410008, China.

Correspondence should be addressed to Lu Xie;luxiex2017@outlook.com

Abstract

In silico T-cell epitope prediction plays an important role in immunization experimental design and vaccine preparation. Currently, most epitope prediction research focuses on peptide processing and presentation, e.g. proteasomal cleavage, transporter associated with antigen processing (TAP) and major histocompatibility complex (MHC) combination. To date, however, the mechanism for immunogenicity of epitopes remains unclear. It is generally agreed upon that T-cell immunogenicity may be influenced by the foreignness, accessibility, molecular weight, molecular structure, molecular conformation, chemical properties and physical properties of target peptides to different degrees. In this work, we tried to combine these factors. Firstly, we collected significant experimental HLA-I T-cell immunogenic peptide data, as well as the potential immunogenic amino acid properties. Several characteristics were extracted, including amino acid physicochemical property of epitope sequence, peptide entropy, eluted ligand likelihood percentile rank (EL rank(%)) score and frequency score for immunogenic peptide. Subsequently, a random forest classifier for T cell immunogenic HLA-I presenting antigen epitopes and neoantigens was constructed. The classification results for the antigen epitopes outperformed the previous research (the optimal AUC=0.81, external validation data set AUC=0.77). As mutational epitopes generated by the coding region contain only the alterations of one or two amino acids, we assume that these characteristics might also be applied to the classification of the endogenic mutational neoepitopes also called ‘neoantigens’. Based on mutation information and sequence related amino acid characteristics, a prediction model of neoantigen was established as well (the optimal AUC=0.78). Further, an easy-to-use web-based tool ‘INeo-Epp’ was developed (available at http://www.biostatistics.online/INeo-Epp/neoantigen.php)for the prediction of human immunogenic antigen epitopes and neoantigen epitopes.

Introduction

An antigen consists of several epitopes, which can be recognized either by B- or T-cells and/or molecules of the host immune system. However, usually only a small number of amino acid residues that comprise a specific epitope are necessary to elicit an immune response [1]. The properties of these amino acid residues causing immunogenicity are unknown. HLA-I antigen peptides are processed and presented as follows: a). cytosolic and nuclear proteins are cleaved to short peptides by intracellular proteinases; b). some are selectively transferred to endoplasmic reticulum (ER) by TAP transporter, and subsequently are treated by endoplasmic reticulum aminopeptidase;c). antigen presenting cells (APCs) present peptides containing 8-11 AA (amino acid) residues on HLA class I molecules to CD8+ T cells [2]. Researchers can now simulate antigen processing and presentation by computational methods to predict binding peptide-MHC complexes (p-MHC). Several types of software systems have been developed, including NetChop [3], NetCTL [4], NetMHCpan [5], MHCflurry [6]. However, the binding to MHC molecules of most peptides is predicted, only 10%~15% of those have been shown to be immunogenic [7-10]. For neoantigens the result was approximately 5% (range, 1%-20%) due to central immunotolerance [11, 12]. As a result, the cycle for vaccine development and immunization research is extended. Here, we aim to develop a T-cell HLA class-I immunogenicity prediction method to further identify real epitopes/neoepitopes from p-MHC to shorten this cycle.

Many experimental human epitopes have been collected and summarized in the immune epitope database (IEDB) [13], which makes it feasible to mathematically predict human epitopes. However there still exist two limitations: i) a high level of MHC polymorphism produces a severe challenge for T-cell epitope prediction. ii) there is an extremely unequal distribution of data to compare epitopes and non-epitopes. It is not conducive to analyze the potential deviation existing in TCR recognition owing to the presentation of different HLA peptides. A general analysis of all HLA presented peptides, ignoring the specific pattern of TCR recognition of individual HLA presented peptides, may result in a lower predictive accuracy.

With the advances in HLA research, Sette *et al* [14] classified, for the first time, overlapping peptide binding repertoires into nine major functional HLA supertypes (A1, A2, A3, A24, B7, B27, B44, B58, B62). In 2008, John Sidney *et al* [15] made a further refinement, in which over 80% of the 945 different HLA-A and -B alleles can be assigned to the original nine supertypes. It has not been reported whether peptides presented by different HLA alleles influence TCR recognition. Hence, we collected experimental epitopes according to HLA alleles and assume that epitopes belonging to the same HLA supertypes have similar properties.

Moreover, screening for endogenic mutational neoepitopes is one of the core steps in tumor immunotherapy. In 2017, Ott PA *et al.* [16]and Sahin *et al* [17]. confirmed that peptides and RNA vaccines made up of neoantigens in melanoma can stimulate and proliferate CD8+ and CD4+ T cells. In addition, a recent research suggests that including neoantigen vaccination not only can expand the existing specific T cells, but also induce a wide range of novel T-cell specificity in cancer patients and enhance tumor suppression[18]. Meanwhile, a tumor can be better controlled by the combination therapy of neoantigen vaccine and programmed cell death protein 1 (PD-1)/PD1 ligand 1(PDL-1) therapy [19, 20]. Nevertheless, a considerable number of predicted candidate p-MHC from somatic cell mutations may be false positive, which would fail to stimulate TCR recognition and immune response. This is undoubtedly a challenge for designing vaccines against neoantigens.

In our study, based on HLA-I T-cell peptides collected from experimentally validated antigen epitopes and neoantigen epitopes, we aim to build a novel method to further reduce the range of immunogenic epitopes screening based on predicted p-MHC. Finally, a simple web-based tool, INeo-Epp (immunogenic epitope/neoepitope prediction), was developed for prediction of human antigen and neoantigen epitopes.

Materials and Methods

The flow chart for 'INeo-Epp' prediction is shown as follows. (see Figure 1)

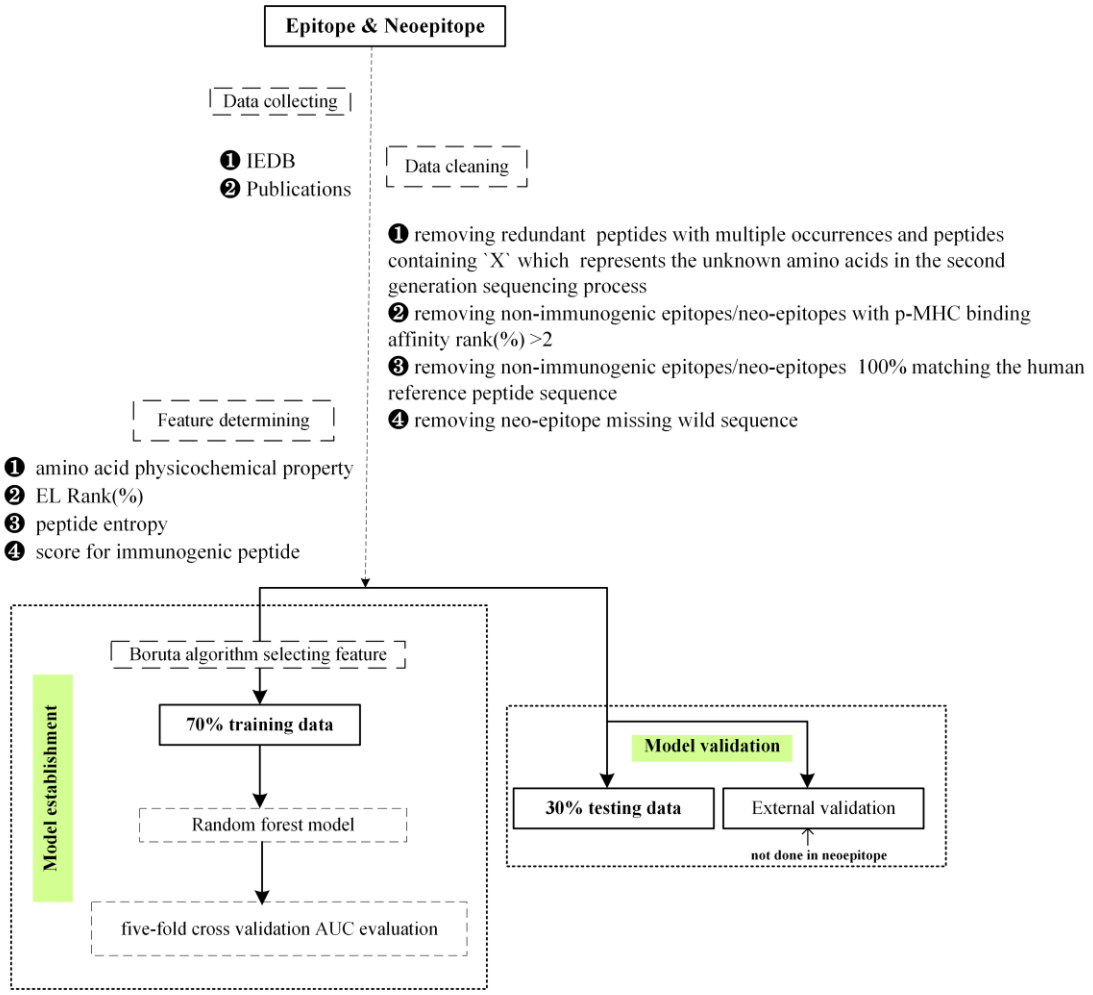

Figure 1: The flow chart for 'INeo-Epp' prediction

Construction of immunogenic and non-immunogenic epitopes

Peptides that can promote cytokine proliferation are considered to be immunogenic epitopes. However, non-immunogenic epitopes may result for the following reasons: a) p-MHC truly unrecognized by TCR; b) peptides not presented by MHC (quantitatively expressed as rank(%)>2, see rank(%) score (below: C24) for details); c) negative selection/clonal presentation induced by excessive similarity to autologous peptides[21]. In this work, to further study the recognition preferences of T cells,

peptides with >2 rank(%) were regarded as not in contact with TCR, and sequences 100% matching the human reference peptides (ftp://ftp.ensembl.org/pub/release-97/fasta/homo_sapiens/pep/) were regarded as exhibiting immune tolerance. Hence, we removed these from the definition of non-immunogenic peptides.

Construction of data sets: epitopes, external validation epitopes and neoepitopes

Antigen epitope data were collected from IEDB (Linear epitope, Human, T cell assays, MHC class I, any disease were chosen). Data collection criteria: each HLA allele quantity >50 and frequency >0.5% (refer to allele frequency database [22]) (Table 1, check Table S1 for detailed information).

Table 1: Summary of IEDB epitope data

HLA supertype	IEDB HLA data	Number		HLA allele frequency Asian / Black / Caucasian	Motif view
		Negative	Positive		
A1	A01:01	811	103	0.154 / 0.046 / 0.164	1-2(ST)-3-4-5-6-7-8-9(Y)
	A26:01	83	19	0.041 / 0.014 / 0.030	1(DE)-2(ITV)-3-4-5-6-7-8-9(FMY)
A2	A02:01	1883	1580	0.049 / 0.123 / 0.275	1-2(LM)-3-4-5-6-7-8-9(ILV)-10(V)
A3	A11:01	196	174	0.139 / 0.014 / 0.060	1-2(IMSTV)-3-4-5-6-7-8-9(K)-10(K)
	A03:01	1400	169	0.063 / 0.083 / 0.139	1-2(ILMTV)-3-4-5-6-7-8-9(K)-10(K)
A24	A24:02	207	219	0.136 / 0.024 / 0.084	1-2(WY)-3-4-5-6-7-8-9(FIW)
	A23:01	1138	12	0.006 / 0.109 / 0.019	1-2(WY)-3-4-5-6-7-8-9-10(F)
B7	B35:01	63	248	0.062 / 0.068 / 0.055	1-2(P)-3-4-5-6-7-8-9(FMY)
	B07:02	523	244	0.034 / 0.005 / 0.0143	1-2(p)-3-4-5-6-7-8-9(FLM)
	B51:01	13	51	0.074 / 0.021 / 0.047	1-2(P)-3-4-5-6-7-8-9(IV)
B8	B08:01	317	195	0.036 / 0.037 / 0.114	1-2-3-4-5(HKR)-6-7-8-9(FILMV)
B27	B27:05	100	86	0.008 / 0.008 / 0.037	1(RY)-2(R)-3(FMLWY)-4-5-6-7-8-9
B44	B37:01	1036	10	0.034 / 0.005 / 0.014	-
	B40:01	67	65	0.022 / 0.012 / 0.052	-
	B44:02	73	66	0.008 / 0.020 / 0.095	1-2(E)-3-4-5-6-7-8-9(FIWY)
B58	B58:01	11	62	0.041 / 0.037 / 0.007	1-2(AST)-3-4-5-6-7-8-9(W)
B62	B15:01	3	70	0.016 / 0.010 / 0.060	1-2(LMQ)-3-4-5-6-7-8-9(FY)
Total		7924	3373		
Remove negative rank(%)>2		5123	3373		
Remove negative human 100% similar		4943	3373		

The external antigen epitope validation set was collected from seven published independent human antigen studies [23-29], consisting of 577 non-immunogenic epitopes and 85 immunogenic epitopes (Table 2, S2 Table)

Table 2: External data included in validation set

Publication time	PMID	Author	non-epitopes	epitopes
2013	23580623	Weiskopf et al	477	42
2018	29397015	Hendrik Luxenburger et al	100	26
2018	30260541	Youchen Xia et al	-	1
2018	30487281	Hawa Vahed et al	-	4
2018	30518652	Atefeh Khakpoor et al	-	2
2018	30587531	Alina Huth et al	-	4
2018	30815394	Solomon Owusu Sekyere et al	-	6
Total			577	85
Remove negative with rank(%) >2 and HLA supertypes (not appeared in training set)			321	69

Here, we removed peptides for which HLA supertypes do not appear in training set, because we assume peptides belonging to the same HLA supertypes to have similar properties. In the external validation set, some peptides bind to rare HLA supertypes. Their characteristics were not included in the training set. Hence, these peptides in the external validation data might lead to a classification bias.

The neoantigens data were collected from 11 publications [19, 30-39] and IEDB mutational epitopes, and 13 published data sets collected by Anne-Mette B in one

publication [40] in 2017 (see Table 3, S3 Table for details) were also included.

Table 3: Neoepitopes data included in this study

Publication time	PMID	Author	Tumor Type	Non-immunogenic neo-epitopes	Immunogenic neo-epitopes	T-cell assay
2013-12	24323902	Darin A. W et al.	Ovarian Cancer	—	1	ELISPOT
2015-9	26359337	Eliezer M et al.	Melanoma	—	18	Clinical benefit
2015-11	26752676	Takahiro K et al.	Lung adenocarcinoma	—	4	—
2016-1	26901407	Alena Gros et al.	Melanoma	12	14	ELISPOT
2016-5	27198675	Erlend Strønen et al.	Melanoma	1134	16	CTL clone
2016-12	28405493	Annika Nelde et al.	Lymphoma	—	2	ELISPOT
2017-6	28619968	Xiuli Zhang et al.	Breast cancer	—	4	Flow cytometry
2017-10	29104575	Markus M et al.	Melanoma	10	16	—
2017-11	29187854	Anne-Mette B et al.	Polytype	1874	42	ELISPOT et al.
2017-11	29132146	Vinod P. B et al.	pancreatic	—	10	Flow Cytometry
2018-5	29720506	Tatsuo Matsuda et al.	Ovarian Cancer	—	3	ELISPOT
2018-12	29409514	Sonntag et al.	pancreatic ductal carcinoma	—	3	Flow Cytometry
2018-10	30357391	Randi Vita et al.	—	6	35	—
Total				3030	168	
Remove duplication				2837	164	
Remove negative rank(%)>2 and human 100% similar				1697	164	

Construction of potential immunogenicity feature

Characteristics calculation of peptides based on amino acid sequences. The formula for calculating peptide characteristics is shown in (1). P_N, P_2, P_C (N-terminal, position 2, C-terminal as anchored sites by default) are considered to be embedded in HLA molecules and no contact with TCRs, therefore not evaluated.

$$\boldsymbol{P_c} = \{\sum_{x\in \boldsymbol{Pos(P)}}^{x\notin \boldsymbol{(N,2,C)}} \boldsymbol{P_{A_c}}\}/(\boldsymbol{len(P)-3}) \quad (1)$$

$\boldsymbol{P}$, peptide. $\boldsymbol{c}$, characteristic. Where $\boldsymbol{P_c}$ represents characteristics of peptides. $\boldsymbol{A}$, amino acid. $\boldsymbol{N}$, N-terminal in a peptide. $\boldsymbol{C}$, C-terminal in a peptide. Pos, amino acid position in peptide. Where $\boldsymbol{P_{Ac}}$ represents characteristics of amino acids in peptides.

Frequency score for immunogenic peptide (C22). Amino acid distribution frequency differences between immunogenicity and non-immunogenic peptides at TCR contact sites (excluding anchor sites) were considered as a feature (2).

$$\boldsymbol{P_{score}} = \sum_{x\in \boldsymbol{Pos(P)}}^{x\notin \boldsymbol{(N,2,C)}} \{\boldsymbol{P_{ie^+}(f'_A)} - \boldsymbol{P_{ie^-}(f'_A)}\} \quad (2)$$

$\boldsymbol{P_{ie}}^+$, immunogenic peptides. $\boldsymbol{P_{ie}}^-$, non-immunogenic peptides. $\boldsymbol{f'_A}$, amino acid frequency in TCR contact position. Where $\boldsymbol{P_{ie+}}$ $\boldsymbol{(f'_A)}$ represents frequency of amino acids in immunogenic peptides at TCR contact sites.

Calculating peptide entropy (C23). Peptide entropy [41] was used as a feature (3).

$$\boldsymbol{P_H} = \{-\sum_{x\in \boldsymbol{Pos(P)}}^{x\notin \boldsymbol{(N,2,C)}} \boldsymbol{P_{f_A}} * \boldsymbol{\log_2(P_{f_A})}\}/(\boldsymbol{len(P)-3}) \quad (3)$$

$\boldsymbol{P_H}$, peptide entropy. $\boldsymbol{f_A}$, amino acid frequency in human reference peptide sequence. Where $\boldsymbol{P_{fA}}$, represents the frequency in human reference peptide sequence of amino acids in epitope peptides.

Rank(%) score (C24). HLA binding prediction were performed using netMHCpan4.0. rank(%) provides a robust filter for the identification of MHC-binding peptides , in which rank(%) was recommended as an evaluation standard，rank(%)<0.5 as strong

154 binders, 0.5<rank(%)<2 as weak binders, rank(%)>2 as no binders.

155 **Five-fold cross-validation, feature selection, random forests and ROC generation.**

156 The 5-fold cross-validation was implemented in R using the package caret [42] (method
157 = "repeatedcv", number = 5, repeats = 3). The feature screening results were generated
158 in R using the package Boruta [43] (a novel random forest based feature selection
159 algorithm for finding all relevant variables, which provides unbiased and stable
160 selection of important and non-important attributes from an information system. It
161 iteratively removes the features which are proven by a statistical test to be less relevant
162 than random probes. It uses Z score (computed by dividing the average loss by its
163 standard deviation) as the importance measure and it takes into account the fluctuations
164 of the mean accuracy loss among trees in the forest). R package randomForest [44] was
165 used for training data (the R language machine learning package caret provides
166 automatic iteration selection of optimal parameters, mtry=15 for antigen epitope,
167 mtry=14 for neoantigen epitope, the remaining parameters use default values). R
168 package ROCR [45] was used for drawing ROC.

169 **Web tool implementation**

170 The front-end of Ineo-Epp was constructed via HTML/JavaScript/CSS. The back end
171 was written in PHP, connecting the web interface and Apache web server. A python
172 script was used for calculating peptide characteristics and extracting mutation
173 information. Models were built using R.

174 **Results**

175 Ultimately, 11,297 validated epitopes and non-epitopes with the length of 8-11 amino
176 acids were collected from IEDB. T-cell responses included activation, cytotoxicity,
177 proliferation, IFN-γ release, TNF release, granzyme B release, IL-2 release, IL-10
178 release, etc. Seventeen different HLA alleles were collected (Fig 2A), and the detailed
179 antigen length distribution is shown in (Fig 2B). Additionally, we collected the
180 neoantigen data from 12 publications, including 2837 non-neoepitopes and 164
181 neoepitopes (Fig 2C), and the detailed neoantigen length distribution is shown in (Fig
182 2D).

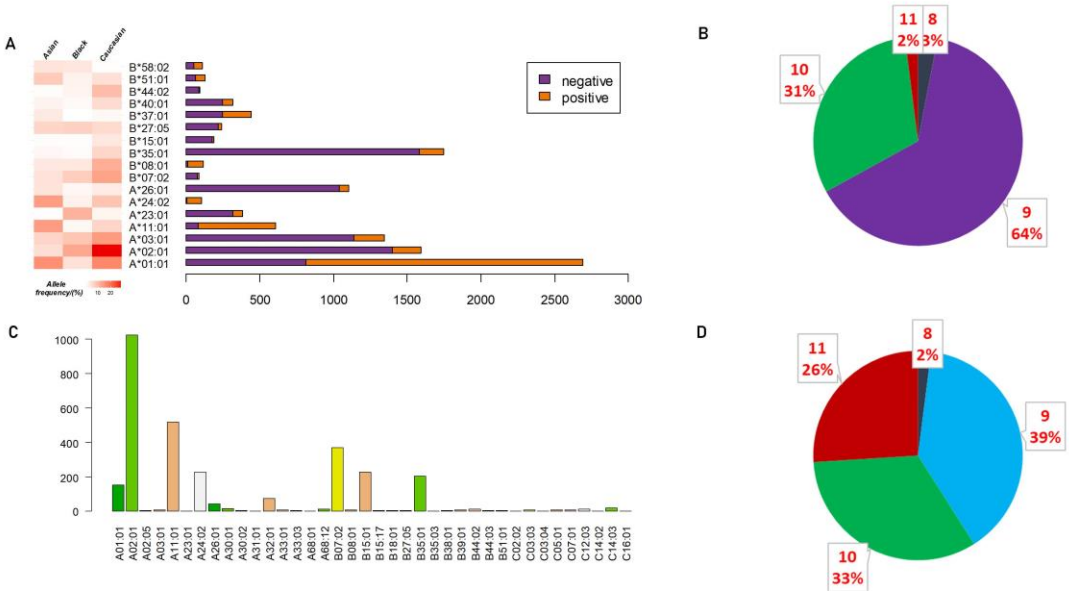

Figure 2: Epitope/neoepitope peptides composition and amino acid lengths distribution. (a) Detailed data distribution of seventeen HLA alleles of antigen peptides and proportion of each HLA allele (positive and negative) epitopes and the corresponding HLA frequency in Asian, Black, Caucasian. (b) Proportion of antigen peptides of 8-11 AA lengths. (c) Data distribution of HLA alleles of neoantigen peptides. (d) Proportion of neoantigen peptides of 8-11 AA lengths.

The TCR contact position plays a crucial role in the analysis of immunogenicity, as TCRs might be more sensitive to some amino acids, the amino acids preference in antigen epitope peptide and antigen non-epitope peptide was further analyzed after excluding anchor sites (N-terminal, position 2, C-terminal) (Fig 3). We found that TCRs tend to identify hydrophobic amino acids. For example, 3/4 hydrophobic amino acids (L, W, P, A, V, M) occur more frequently in immunogenicity epitopes. Charged amino acids (*e.g.* D, K) are enriched in non-epitopes whereas the rest of charged amino acids (R, H, E) show no difference .Based on the result in figure 3, the amino acid distribution difference at the TCR contact sites was regarded by us as one of the immunogenicity features (*i.e.* Frequency score for immunogenic peptide (C22)).

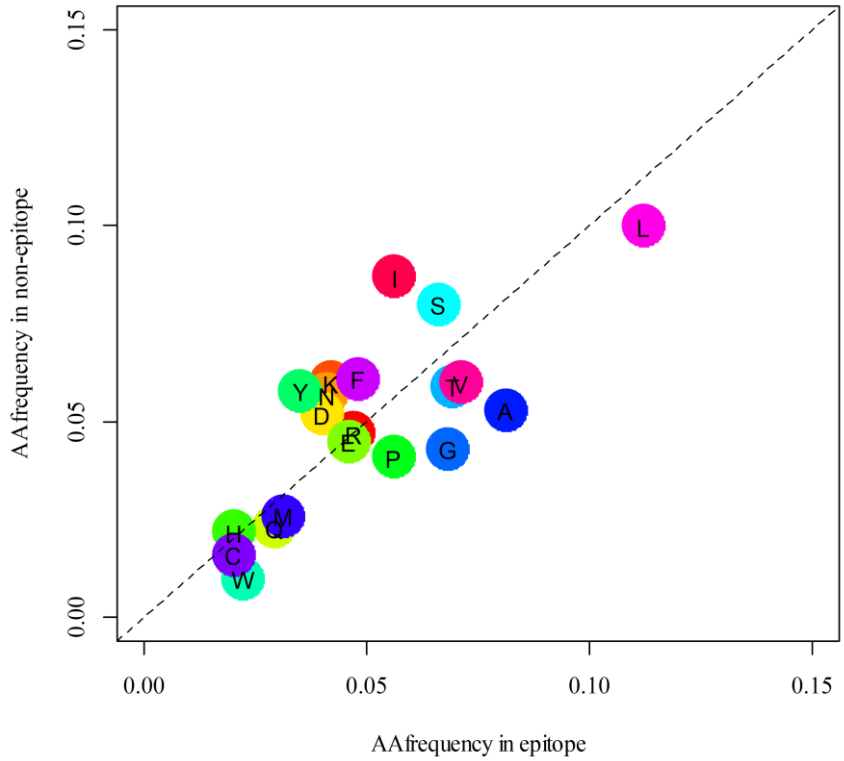

Figure 3: Antigen epitope amino acid distribution frequency in TCR contact site of epitopes and non-epitopes. Frequency distribution of amino acids at TCR contact sites in antigen epitope and non-epitope peptides, and the amino acids below the dotted line are preferred by the epitope.

Classification prediction model for antigen epitopes

We constructed the features of peptides on the basis of the characteristics of amino acids (see Materials and Methods section: Characteristics Calculation of peptides based on amino acids). All amino acid characteristics were selected from Protscale [46] in ExPASy (SIB bioinformatics resource portal). The 21 involved features are as follows: Kyte–Doolittle numeric hydrophobicity scale (C1) [47], molecular weight (C2), bulkiness (C3) [48], polarity (C4) [49], recognition factors (C5) [50], hydrophobicity (C6) [51]，retention coefficient in HPLC (C7) [52] , ratio hetero end/side (C8)[49], average flexibility (C9) [53], beta-sheet (C10) [54], alpha-helix (C11) [55]，beta-turn (C12) [55]，relative mutability (C13) [56], number of codon(s) (C14), refractivity (C15) [57], transmembrane tendency (C16) [58]，accessible residues (%) (C17) [59]，average area buried (C18) [60]，conformational parameter for coil (C19) [55]， total beta-strand (C20) [60]， parallel beta-strand (C21) [61] (see Table S4 in detail). Also, frequency score for immunogenic peptide (C22), peptide entropy (C23) and rank(%) (C24) were also taken into consideration. Together, 24 immunogenic features were collected, and all features were retained for antigen epitopes prediction after screening using the R package Boruta. Compared with other characteristics, the frequency score for

immunogenic peptide and rank(%) have higher impact, suggesting they have more significant influence on antigen epitopes classification (Figure 4A).

The receiver operator characteristic (ROC) curve of models are shown in Fig 4. The five-fold cross validation AUC was 0.81 in the prediction model for antigen epitope (line in red Fig 4B) and the externally validated (see table 2) AUC was 0.75 (line in purple Fig 4C). Here, we tried to remove peptides for which HLA supertypes not appearing in training set from the externally validated antigen data and, the AUC, specificity, and sensitivity were increased to 0.78, 0.71, and 0.72, respectively. (line in pink Fig4 C). This, to some extent, verifies our conjecture about TCR specific recognition of different HLA alleles presenting peptides.

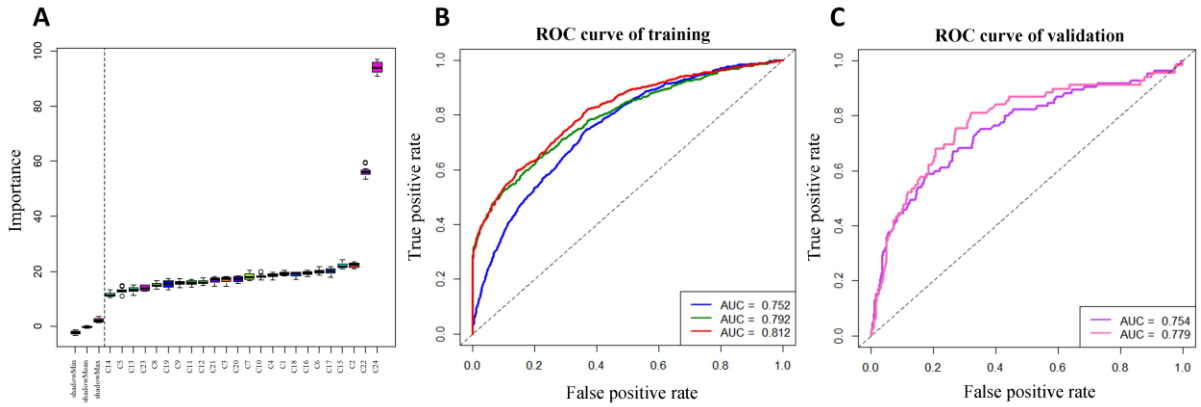

Figure 4: Feature selection in antigen epitopes and ROC curves of antigen epitopes classification. (a)Peptide features: Twenty four features were screened and we defined the features on the right of the dotted line as being effective. (b)Trained model: The line in blue represents antigen epitopes without screening; the line in green represents selection with the deletion of rank(%)>2 non-epitope; and the line in red represents selection with the deletion of the non-epitopes 100% matching human reference peptide sequence. (c)External validation: The ROC curves for the external verification set, line in purple represents modeling using antigen epitopes without filtering, the line in pink represents modeling using antigen epitopes removing non-epitopes which rank(%)>2 and HLA for which supertypes not appearing in training set.

Classification prediction model for neoantigen epitopes

Neoantigens derived from somatic mutations are different from the wild peptide sequences. Therefore, some mutation-related characteristics were also taken into account. For instance, difference in hydrophobility before and after mutation (C25), differential agretopicity index (DAI, C26) [62] and whether the mutation position was anchored (C27). Finally, 27 features were selected for the neoantigen epitope prediction model. However, only 25 neoantigen related features were retained after running Boruta, because C25 and C27 were removed. Also, rank(%) showed a marked effect (Fig 5A). in the five-fold cross-validation of the prediction model for neoantigen epitopes, AUC was 0.78 (Fig 5B).

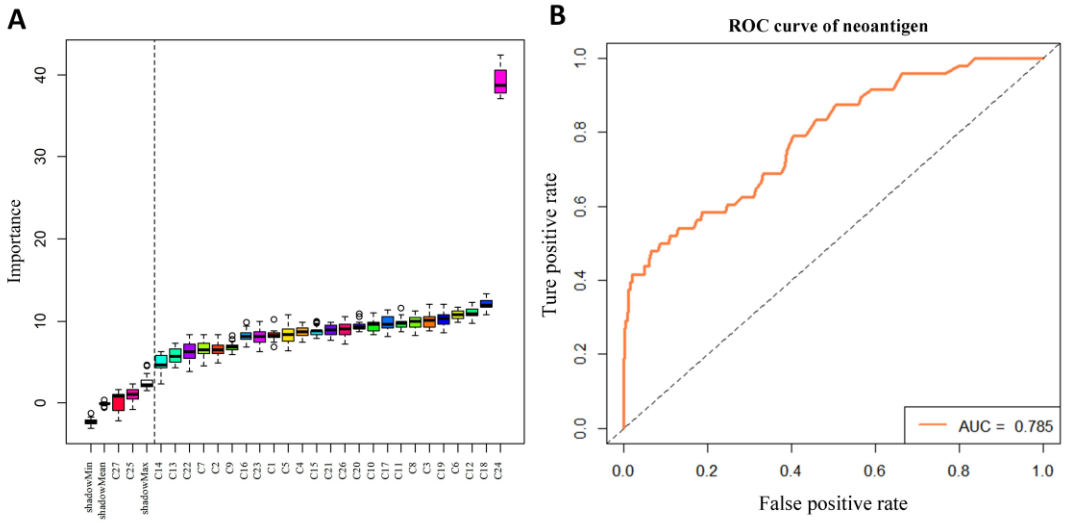

Figure 5: Feature selection in neoantigen epitopes and ROC curves of neoantigen epitopes classification. (a) Twenty seven features were screened and the 25 features on the right of the dotted line were reserved for modeling using a random forest algorithm. (b) ROC curves of neoantigen epitopes classification.

Web server for TCR epitope prediction

Based on these above-mentioned validated features, we established a web server for TCR epitope prediction, named 'INeo-Epp'. This tool can be used to predict both immunogenic antigen and neoantigen epitopes. For antigen, the nine main HLA supertypes can be used. We recommend the peptides with the lengths of 8-12 residues, but not less than 8. N-terminal, position 2, C-terminal were treated as anchored sites by default. A predictive score value greater than 0.5 is considered as immunogenicity (Positive-High),the score between 0.4-0.5 is considered as (Positive-Low),the score less than 0.4 is considered as (Negative-High).It is critical to make sure that HLA-subtype must match your peptides(rank(%)<2). Where HLA-subtypes mismatch, the large deviation of rank(%) value may strongly influence the results. Additionally, the neoantigen model requires providing wild type and mutated sequences at the same time to extract mutation associated characteristics, and currently only immunogenicity prediction for neoantigens of single amino acid mutations are supported. Users can choose example options to test the INeo-Epp (http://www.biostatistics.online/INeo-Epp/neoantigen.php).

Discussion

Due to the complexity of antigen presenting and TCR binding, the mechanism of TCR recognition has not been clearly revealed. In 2013, J. A. Calis [63] developed a tool for epitope identification for mice and humans (AUC = 0.68). Although mice and human beings are highly homologous, the murine epitopes may very likely cause limitations in identifying human epitopes. Inspired by J. A. Calis , our research here focused on human beings' epitopes and has been conducted in a larger data set.

By analyzing epitope immunogenicity from the perspective of amino acid molecular composition, we observed that TCRs do have a preference for hydrophobic amino acid recognition. For short peptides presented by different HLA supertypes, TCRs may have different identification patterns. The immunogenicity prediction based on all HLA-presenting peptides may affect the accuracy of the prediction results. That is, if the prediction could focus on specified HLA-presenting peptides the results may improve. Therefore in our work we used HLA supertypes to improve the prediction of HLA-presenting epitopes, including antigen epitopes and neoantigen epitopes, for a better recognition by TCRs. At present, neoantigen epitopes that can be collected in accordance with the standard for experimental verification are too few, the data of positive and negative neoantigens are unbalanced, and there is not enough data to be used for external verification set. In the future, we will continue to refine and expand our training and verification datasets. Recently, Céline M. Laumont [64] demonstrated that noncoding regions aberrantly expressing tumor-specific antigens (aeTSAs) may represent ideal targets for cancer immunotherapy. These epitopes can also be studied in the future. Increased epitope data may also help empower the prediction of potentially immunogenic peptides or neopeptides.

Conclusions

Neoantigen prediction is the most important step at the start of preparation of neoantigen vaccine. Bioinformatics methods can be used to extract tumor mutant peptides and predict neoantigens. Most current strategies aimed at ended in presenting peptides predictions and among the results of these predictions, probably only fewer than 10 neoantigens might be clinically immunogenic and produce effective immune response. It is time-consuming and costly to experimentally eliminate the false positively predicted peptides. Our methods as developed in this study and the INeo-Epp tool may help eliminate false positive antigen/neoantigen peptides, and greatly reduce the amount of candidates to be verified by experiments. We believe that in the age of biological systems data explosion, computational approaches are a good way to enhance research efficiency and direct biological experiments. With the development of machine learning and deep learning, we expect the prediction of epitope immunogenicity will be continually improved.

In summary, this study provides a novel T-cell HLA class-I immunogenicity prediction method from epitopes to neoantigens, and the INeo-Epp can be applied not only to identify putative antigens, but also to identify putative neoantigens.

It needs to be stated here that we published the preprint [65] of this article in July 2019.This is a modified version.

Data Availability

The data used to support the findings of this study are included within the supplementary information file(s).

Competing of Interests

The author(s) declare(s) that there is no conflict of interest regarding the publication of this paper

Funding Statement

This work was funded by the National Natural Science Foundation of China (No. 31870829), Shanghai Municipal Health Commission, and Collaborative Innovation Cluster Project (No. 2019CXJQ02). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgments

We sincerely thank Drs. Menghuan Zhang, Hong Li and Qibing Leng for valuable discussion. We also acknowledge Dr. Michael Liebman for his critical reading and editing.

Supplementary Material

S1 Table **IEDB antigen epitopes summary.** Detailed description of 17 HLA molecules collected from IEDB. (XLSX)
S2 Table **External validation antigen epitopes summary.** Epitope details of 7 publications. (XLSX)
S3 Table **Neoantigen epitopes summary.** Epitope details of 13 publications. (XLSX)
S4 Table **Summary of amino acid characteristics.** For all amino acid characteristics (n=21) that are described in the ExPASy. (XLSX)

References

[1] D. V. Desai, and U. Kulkarni-Kale, "T-cell epitope prediction methods: an overview," Methods Mol Biol, vol. 1184, pp. 333-64, 2014.
[2] A. L. Goldberg, and K. L. Rock, "Proteolysis, proteasomes and antigen presentation," *Nature,* vol. 357, no. 6377, pp. 375-379,1992.
[3] K. Can, A. K. Nussbaum, S. Hansjörg *et al.*, "Prediction of proteasome cleavage motifs by neural networks," *Protein Eng*, no. 4, pp. 4, 2002.
[4] M. V. Larsen, C. Lundegaard, K. Lamberth *et al.*, "An integrative approach to CTL epitope prediction: A combined algorithm integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions," *European Journal of Immunology,* vol. 35, no. 8, pp. 2295-2303,2005.
[5] V. Jurtz, S. Paul, M. Andreatta *et al.*, "NetMHCpan-4.0: Improved Peptide–MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data," *Journal of Immunology,* vol. 199, no. 9, pp. ji1700893, 2017.
[6] T. J. O'Donnell, A. Rubinsteyn, M. Bonsack et al., "MHCflurry: Open-Source Class I MHC Binding Affinity Prediction," Cell Syst, vol. 7, no. 1, pp. 129-132.e4, Jul 25, 2018.

[7] M. Wang, K. Lamberth, M. Harndahl *et al.*, "CTL epitopes for influenza A including the H5N1 bird flu; genome-, pathogen-, and HLA-wide screening," *Vaccine,* vol. 25, no. 15, pp. 0-2831.2007.

[8] C. L. Perez, M. V. Larsen, R. Gustafsson *et al.*, "Broadly Immunogenic HLA Class I Supertype-Restricted Elite CTL Epitopes Recognized in a Diverse Population Infected with Different HIV-1 Subtypes," *Journal of Immunology,* vol. 180, no. 7, pp. 5092-5100,2008.

[9] C. Lundegaard, I. Hoof, O. Lund *et al.*, "State of the art and challenges in sequence based T-cell epitope prediction," *Immunome Research,* vol. 6 Suppl 2, no. Suppl 2, pp. S3, 2010.

[10] J. L. Sanchez-Trincado, G.-P. Marta, and R. P. A., "Fundamentals and Methods for T- and B-Cell Epitope Prediction," *Journal of Immunology Research,* vol. pp. 1-14,2017.

[11] E. G. Phimister, and V. N. Kristensen, "The Antigenicity of the Tumor Cell — Context Matters," *New England Journal of Medicine,* vol. 376, no. 5, pp. 491-493,2017.

[12] K. Kiyotani, H. T. Chan, and Y. Nakamura, "Immunopharmacogenomics towards personalized cancer immunotherapy targeting neoantigens," Cancer Sci, vol. 109, no. 3, pp. 542-549, Mar, 2018.

[13] V. Randi, J. A. Overton, J. A. Greenbaum *et al.*, "The immune epitope database (IEDB) 3.0," *Nucleic Acids Research*, no. D1, pp. D1, 2014.

[14] A. Sette, and J. Sidney, "Nine major HLA class I supertypes account for the vast preponderance of HLA-A and -B polymorphism," Immunogenetics, vol. 50, no. 3-4, pp. 201-12, Nov, 1999.

[15] J. Sidney, B. Peters, N. Frahm *et al.*, "HLA class I supertypes: a revised and updated classification," vol. 9, no. 1, pp. 1-0,2008.

[16] "An immunogenic personal neoantigen vaccine for patients with melanoma."

[17] "Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer," *Nature,* vol. 547, no. 7662, pp. 222-226,2017.

[18] Z. Hu, P. A. Ott, and C. J. Wu, "Towards personalized, tumour-specific, therapeutic vaccines for cancer," Nat Rev Immunol, vol. 18, no. 3, pp. 168-182, Mar, 2018.

[19] E. M. Van Allen, D. Miao, B. Schilling *et al.*, "Genomic correlates of response to CTLA-4 blockade in metastatic melanoma," *Science,* vol. 350, no. 6257, pp. 207-211, 2015.

[20] M. Efremova, F. Finotello, D. Rieder et al., "Neoantigens Generated by Individual Mutations and Their Role in Cancer Immunity and Immunotherapy," Front Immunol, vol. 8, pp. 1679, 2017.

[21] L. Klein, M. Hinterberger, G. Wirnsberger *et al.*, "Antigen presentation in the thymus for positive selection and central tolerance induction," *Nature reviews. Immunology,* vol. 9, no. 12, pp. 833-844,2009.

[22] F. F. Gonzalez-Galarza, A. McCabe, E. J. Melo Dos Santos et al., "Allele

Frequency Net Database," Methods Mol Biol, vol. 1802, pp. 49-62, 2018.

[23] D. Weiskopf, M. A. Angelo, E. L. D. Azeredo *et al.*, "Comprehensive analysis of dengue virus-specific responses supports an HLA-linked protective role for CD8(+) T cells," *Proc Natl Acad Sci U S A,* vol. 110, no. 22, pp. E2046-E2053, 2013.

[24] H. Luxenburger, F. Grass, J. Baermann et al., "Differential virus-specific CD8(+) T-cell epitope repertoire in hepatitis C virus genotype 1 versus 4," J Viral Hepat, vol. 25, no. 7, pp. 779-790, Jul, 2018.

[25] Y. Xia, W. Pan, X. Ke *et al.*, "Differential escape of HCV from CD8+ T cell selection pressure between China and Germany depends on the presenting HLA class I molecule," *Journal of Viral Hepatitis,* vol. 26, no. 1, pp. 73-82, 2019.

[26] H. Vahed, A. Agrawal, R. Srivastava *et al.*, "Unique Type I Interferon, Expansion/Survival Cytokines, and JAK/STAT Gene Signatures of Multifunctional Herpes Simplex Virus-Specific Effector Memory CD8 T Cells Are Associated with Asymptomatic Herpes in Humans," *Journal of Virology,* vol. 93, no. 4, pp. e01882-18, 2019.

[27] A. Khakpoor, Y. Ni, A. Chen et al., "Spatiotemporal Differences in Presentation of CD8 T Cell Epitopes during Hepatitis B Virus Infection," J Virol, vol. 93, no. 4, Feb 15, 2019.

[28] A. Huth, X. Liang, S. Krebs et al., "Antigen-Specific TCR Signatures of Cytomegalovirus Infection," J Immunol, vol. 202, no. 3, pp. 979-990, Feb 1, 2019.

[29] S. O. Sekyere, B. Schlevogt, F. Mettke *et al.*, "HCC immune surveillance and antiviral therapy of hepatitis C virus infection," *Liver cancer,* vol. 8, no. 1, pp. 41-65, 2019.

[30] D. A. Wick, J. R. Webb, J. S. Nielsen *et al.*, "Surveillance of the Tumor Mutanome by T Cells during Progression from Primary to Recurrent Ovarian Cancer," *Clinical Cancer Research,* vol. 20, no. 5, 2013.

[31] T. Karasaki, K. Nagayama, M. Kawashima *et al.*, "Identification of Individual Cancer-Specific Somatic Mutations for Neoantigen-Based Immunotherapy of Lung Cancer," *Journal of Thoracic Oncology Official Publication of the International Association for the Study of Lung Cancer,* vol. 11, no. 3, pp. 324-333, 2015.

[32] A. Gros, M. R. Parkhurst, E. Tran *et al.*, "Prospective identification of neoantigen-specific lymphocytes in the peripheral blood of melanoma patients," *Nature Medicine,* vol. 22, no. 4, pp. 433-438,2016.

[33] E. Strønen, M. Toebes, S. Kelderman *et al.*, "Targeting of cancer neoantigens with donor-derived T cell receptor repertoires," *Science,* vol. 352, no. 6291, pp. 1337-1341, 2016.

[34] A. Nelde, J. S. Walz, D. J. Kowalewski *et al.*, "HLA class I-restricted MYD88 L265P-derived peptides as specific targets for lymphoma immunotherapy," *OncoImmunology,* vol. 6, no. 3, Mar 4, 2017.

[35] X. Zhang, S. Kim, J. Hundal *et al.*, "Breast Cancer Neoantigens Can Induce CD8 T-Cell Responses and Antitumor Immunity," *Cancer Immunology Research,* vol. 5, no. 7, pp. 516-523, 2017.

[36] M. Markus, G. David, C. George *et al.*, "'Hotspots' of Antigen Presentation Revealed by Human Leukocyte Antigen Ligandomics for Neoantigen Prioritization," *Front Immunol,* vol. 8, pp. 1367,2017

[37] V. P. Balachandran, M. Łuksza, J. N. Zhao *et al.*, "Identification of unique neoantigen qualities in long-term survivors of pancreatic cancer," *Nature,* vol. 551, no. 7681, pp. 512-516,2017.

[38] T. Matsuda, M. Leisegang, J.-H. Park *et al.*, "Induction of Neoantigen-Specific Cytotoxic T Cells and Construction of T-cell Receptor-Engineered T Cells for Ovarian Cancer," *Clinical cancer research : an official journal of the American Association for Cancer Research,* vol. 24, no. 21, pp. 5357-5367, 2018.

[39] K. Sonntag, H. Hashimoto, M. Eyrich *et al.*, "Immune monitoring and TCR sequencing of CD4 T cells in a long term responsive patient with metastasized pancreatic ductal carcinoma treated with individualized, neoepitope-derived multipeptide vaccines: a case report," *Journal of translational medicine,* 16,2018.

[40] A.-M. Bjerregaard, M. Nielsen, V. Jurtz *et al.*, "An Analysis of Natural T Cell Responses to Predicted Tumor Neoepitopes," *Frontiers in immunology,* 8, 2017.

[41] C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal,* vol. 27, 1948.

[42] M. Kuhn, "Building Predictive Models in R Using the caret Package," *Journal of Statistical Software*, 2008.

[43] M. B. Kursa, and W. R. Rudnicki, "Feature Selection with the Boruta Package," *Journal of Statistical Software,* vol. 036, 2010.

[44] A. Liaw, and M. Wiener, "Classification and Regression by randomForest," *R News,* vol. 23, no. 23, 2002.

[45] T. Sing, O. Sander, N. Beerenwinkel *et al.*, "ROCR: visualizing classifier performance in R," *Bioinformatics (Oxford, England),* vol. 21, no. 20, pp. 3940-3941, 2005.

[46] Walker, and M. J., "The proteomics protocols handbook," *Biochemistry,* vol. 71, no. 6, pp. 696-696, 2006.

[47] J. Kyte, and R. F. Doolittle, "A simple method for displaying the hydropathic character of a protein," vol. 157, no. 1, pp. 105-132,1982.

[48] J. M. Zimmerman, N. Eliezer, and R. Simha, "The characterization of amino acid sequences in proteins by statistical methods," *Journal of theoretical biology,* vol. 21, no. 2, pp. 170-201,1968.

[49] Grantham, and R., "Amino Acid Difference Formula to Help Explain Protein Evolution," *Science,* vol. 185, no. 4154, pp. 862-864,1974.

[50] Fraga, and Serafin, "Theoretical prediction of protein antigenic determinants from amino acid sequences," *Canadian Journal of Chemistry,* vol. 60, no. 20,

pp. 2606-2610,1982.

[51] R. M. Sweet, and D. Eisenberg, "Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure," *Journal of molecular biology,* vol. 171, no. 4, pp. 479-488,1983.

[52] Meek, and L. J., "Prediction of peptide retention times in high-pressure liquid chromatography on the basis of amino acid composition," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 77, no. 3, pp. 1632-1636,1980.

[53] G. D. Rose, A. R. Geselowitz, G. J. Lesser *et al.*, "Hydrophobicity of amino acid residues in globular proteins," *Science (New York, N.Y.),* vol. 229, no. 4716, pp. 834-838, 1985.

[54] P. Y. Chou, and G. D. Fasman, "Prediction of the secondary structure of proteins from their amino acid sequence," *Advances in enzymology and related areas of molecular biology,* vol. 47, pp. 45-148, 1978, 1978.

[55] G. Deléage, and B. Roux, "An algorithm for protein secondary structure prediction based on class prediction," *Protein engineering,* vol. 1, no. 4, pp. 289-294, 1987 Aug-Sep, 1987.

[56] A. Burger, "Atlas of Protein Sequence and Structure 1969," *Journal of Medicinal Chemistry,* vol. 13, no. 2, pp. 337-337, 1970.

[57] D. D. Jones, "Amino acid properties and side-chain orientation in proteins: A cross correlation approach," *Journal of Theoretical Biology,* vol. 50, no. 1, pp. 167-183,1975.

[58] G. Zhao, and E. London, "Strong Correlation Between Statistical Transmembrane Tendency and Experimental Hydrophobicity Scales for Identification of Transmembrane Helices," *Journal of Membrane Biology,* vol. 229, no. 3, pp. p.165-168,2009.

[59] J. Janin, "Surface and inside volumes in globular proteins," *Nature,* vol. 277, no. 5696, pp. 491-492, 1979.

[60] J. R. Green, M. J. Korenberg, R. David *et al.*, "Recognition of Adenosine Triphosphate Binding Sites Using Parallel Cascade System Identification," *Annals of Biomedical Engineering,* vol. 31, no. 4, pp. 462-470,2003.

[61] S. Lifson, and C. Sander, "Antiparallel and parallel beta-strands differ in amino acid residue preferences," *Nature,* vol. 282, no. 5734, pp. 109-111, 1979.

[62] F. Duan, J. Duitama, S. Al Seesi *et al.*, "Genomic and bioinformatic profiling of mutational neoepitopes reveals new rules to predict anticancer immunogenicity," *J Exp Med,* vol. 211, no. 11, pp. 2231-48, Oct 20, 2014.

[63] J. J. A. Calis, M. Maybeno, J. A. Greenbaum *et al.*, "Properties of MHC class I presented peptides that enhance immunogenicity," *PLoS computational biology,* vol. 9, no. 10, pp. e1003266,, 2013.

[64] C. M. Laumont, K. Vincent, L. Hesnard *et al.*, "Noncoding regions are the main source of targetable tumor-specific antigens," *Sci Transl Med,* vol. 10, no. 470, Dec 5, 2018.

[65] G. Wang, H. Wan, X. Jian *et al.*, "INeo-Epp: T-cell HLA class I immunogenic or neoantigenic epitope prediction via random forest algorithm based on sequence related amino acid features," bioRxiv, 2019.