Hardianto & Yusuf (2019) Vol. 7 No. 2: 51-62

Chimica et Natura Acta

p-ISSN: 2355-0864 e-ISSN: 2541-2574

Homepage: http://jurnal.unpad.ac.id/jcena

Development of Predictive Model for Helper T Lymphocyte Epitope Binding to HLA-DRB1*01:01

Ari Hardianto*, Muhammad Yusuf

Department of Chemistry, Faculty Mathematics and Natural Sciences, Universitas Padjadjaran, Jln. Raya Bandung-Sumedang km. 21 Jatinangor, Sumedang, West Java, 45363, Indonesia
*Corresponding author: a.hardianto@unpad.ac.id

DOI: https://doi.org/10.24198/cna.v7.n2.23713

Abstract: Epitopes are essential peptides for immune system stimulation, such as governing helper T lymphocyte (HTL) activation via antigen presentation and recognition. Current predictive models for epitope selection mainly rely on the antigen presentation, although HTLs only recognize 50% of the presented peptides. Thus, we developed a HTL epitope predictor which involves the antigen recognition step. The predictor is specific for epitopes presented by Human Leukocyte Allele (HLA)-DRB1*01:01, which is protective against developing multiple sclerosis and association with autoimmune diseases. As the data set, we used binding register of immunogenic and non-immunogenic HTL peptides related to HLA-DRB1*01:01. The binding registers were obtained from consensus results of two current HLA-binder predictors. Amino acid descriptors were extracted from the binding registers and subjected to random forest algorithm. A threshold optimization were applied to overcome data set imbalance class. In addition, descriptors were screened by using a recursive feature elimination to enhance the model performance. The obtained model shows that the hydrophobicity, steric, and electrostatic properties of epitopes, mainly at center of binding registers, are important for the TCR recognition as well as the HTL epitopes predictive model. The model complements current HLA-DRB1*01:01-binder prediction methods to screen immunogenic HTL epitopes.

Keywords: epitope, predictive model, HLA-DRB1*01:01, helper T lymphocyte, random forest algorithm

***Abstrak*:** *Epitop adalah peptida yang sangat penting dalam stimulasi sistem kekebalan, seperti dalam pengaturan aktivasi limfosit T penolong (HTL) melalui presentasi dan pengenalan antigen. Saat ini, model prediksi untuk menyeleksi epitop hanya berdasarkan pada presentasi antigen, meskipun HTL hanya mengenali 50% peptida yang dipresentasikan. Maka, kami mengembangkan model prediksi yang melibatkan tahapan pengenalan antigen. Model prediksi yang dikembangkan spesifik untuk epitop yang dipresentasikan oleh Human Leukocyte Allele (HLA)-DRB1*01:01, yang bersifat proteksi terhadap sklerosis ganda dan berkaitan dengan penyakit-penyakit autoimun. Sisi pengikatan peptida HTL yang imunogenik dan non-imunogenik pada HLA-DRB1*01:01 digunakan sebagai data set. Informasi sisi pengikatan diperoleh dari hasil konsensus dua server prediksi peptida. Selanjuntya, deskriptor asam amino diekstrak dari sisi pengikatan peptida dan digunakan untuk melatih model algoritma random forest. Pendekatan optimasi ambang juga digunakan untuk mengatasi ketidakseimbangan jumlah kelas pada data set. Selain itu, deskriptor diseleksi dengan metode eliminasi rekursif untuk meningkatkan performa model. Model yang dihasilkan menunjukkan bahwa hidrofobisitas, sterik, dan elektrostatik epitop, terutama pada bagian sisi pengikatan peptida ke MHC, penting bagi pengenalan TCR. Model prediksi ini melengkapi metode prediksi peptida yang terikat pada HLA-DRB1*01:01 untuk menyeleksi epitop HTL yang imunogenik.*

Kata kunci: *epitop, model prediktif, helper T lymphocyte, algoritma random forest*

INTRODUCTION

In human, the adaptive immune system has an essential role in protecting hosts from diverse pathogen invasions. Through CD8+ cytotoxic T lymphocytes (CTLs), it destroys either infected or tumor cells. Meanwhile, activating CD4+ helper T lymphocytes (HTL) provokes other immune cells like B cells and macrophages to eventually destroy pathogens (Murphy 2011). These responses are relied on two important molecular events. The first event is the antigen presentation, where the antigenic peptide epitope binds to the major histocompatibility complex (MHC) as pMHC. The second event is the T cell receptor (TR) recognition of pMHC which results in the activation of either CTL or HTL (Khan & Ranganathan 2011).

CTLs and HTLs recognize dissimilar peptide epitopes displayed by two different MHC molecules (Murphy 2011). MHC class I molecules load peptide epitopes (CTL epitopes) for CTL recognition, whereas MHC class II molecules present other kinds of peptide epitopes (HTL epitopes) for HTL. CTL epitopes, possessing length 9-11 residues, are intracellular pathogens origin. They particularly bind the MHC class I through their N- and C-termini residues, as the consequence their middle parts have a bulged conformation. On the other hand, HTL epitopes are generated through a serial antigen processing of extracellular pathogens and have length 12-25 residues. They use their nine sequential amino acids to bind MHC class II. These nine residues, which are called peptide binding registers, mainly interact with the MHC class II at positions 1, 4, 6, and 9 (Sant'Angelo *et al*. 2002). The rest residues of HTL epitopes, referred as peptide flanking residues or PFRs, can extend outside N- and C-termini of the groove.

Peptide epitope sequences are vital information to develop vaccines in prophylactic and immunotherapeutic settings. Such information assisted the development of the next-generation Malaria RTS,S vaccine (MosquirixTM) which acquired a final recommendation from WHO in 2015 (Oyarzún & Kobe 2016). Interestingly, peptide epitope information also contributed to the development of the RNA-lipoplex vaccine targeting melanoma which reached a phase I clinical trial in 2016 (Gilboa 2016).

Unfortunately, peptide epitope discovery through experimental methods are laborious, time consuming, and costly (Tong *et al*. 2007). To address such problems, many research groups worldwide have developed computational approaches (i.e. immunoinformatics') either using structure- (Rognan *et al*. 1994; Rosenfeld *et al*. 1995; Tong *et al*. 2004; Bordner & Abagyan 2006; Khan & Ranganathan 2010; Patronov *et al*. 2011) or sequence-based approaches (Rammensee *et al*. 1999; Guan *et al*. 2003; Reche *et al*. 2004; Zhang *et al*. 2005; Nielsen *et al*. 2004; Gonzalez-Galarza *et al*. 2011; Zhang *et al*. 2012; Karosiene *et al*. 2013; Andreatta *et al*. 2015). These two methods complement each other, where the structural approach needs sequences and vice versa. Both approaches construct their algorithms based on the binding of the peptide epitope to the MHC, because this antigen presentation event is considered as the critical step in the immune system activation. Most methods assume that the higher binding affinity values of MHC-bound peptides are, the longer time they are presented. Hence, they have a bigger change to be recognized by T cell. Studies revealed that only a half of such peptide MHC-binders are recognized by T cells or immunogenic (Chuan & Ranganathan 2013), however.

In last a decade, some groups started to develop immunogenicity prediction methods for CTL epitopes related to HLA-A2. The first predictor was POPI. It used 23 physicochemical properties and feed them to a super vector machine (SVM) classifier. The same group then developed a POPISK using SVM with string kernels (Tung *et al*. 2011). It outperformed POPI and identified six important positions (1, 4, 5, 6, 8, and 9) for CTL epitopes immunogenicity. Another group, Saethang *et al*. (2013), built a PAAQD using a random forest based on amino acid pairwise contact potentials (AAPP) and quantum topological molecular similarity (QTMS) descriptors. They suggested that the positions 1 and 8 determine the immunogenicity of nonamer peptide epitopes, whereas the anchor residues less contribute in T-cell reactivity prediction. Chowell *et al*. (2015) analyzed a hydrophobicity difference between immunogenic and non-immunogenic CTL peptides. They found that the hydrophobicity property is sufficient to predict immunogenic CTL epitopes. Zhang *et al*. (2015) applied genetic algorithm-based ensemble learning, as a feature selection, on various combination of physicochemical descriptors. They proposed that relative accessible surface areas (RASA) of peptides and AAPP are the optimal features for CTL epitopes immunogenicity.

On the other hand, none of method for HTL epitopes immunogenicity is available yet. The complexity of HTL epitopes, which consist of binding registers and PFRs, is the major challenge in their immunogenicity modeling. Intriguingly, these peptide epitopes mainly interact with TRs through their binding registers (Sant'Angelo *et al*. 2002). This basis could be sufficient for discriminating immunogenic HTL epitopes from MHC class II-binder peptides predicted by current prediction servers.

The HLA-DRB1*01:01 is a kind of human MHC class II allele. It is protective against developing multiple sclerosis and association with autoimmune diseases, for example rheumatoid arthritis (Sauer *et al*. 2015) and Crohn's disease (Goyette *et al*. 2015). Incidence and prevalence of these autoimmune diseases worldwide is increasing (Lerner *et al*. 2015). During 2000 to 2016, the number of 583,694 people worldwide were suffering multiple sclerosis, and 108,907 of them reside in the lower middle-income region, including Indonesia. Furthermore, in the same period of time, around 5 million people in the world live with rheumatoid arthritis (WHO 2016). Thus, modelling immunogenic HTL epitopes related to HLA-DRB1*01:01 is important.

Despite limited experimental data providing binding register information of HTL epitopes, the NetMHCIIpan (Andreatta *et al*. 2015) and the TEPITOPEpan (Zhang *et al*. 2012) exhibited a reliable prediction of human MHC class II-bound peptides, including the HLA-DRB1*01:01 (Andreatta

et al. 2015). Hence, we constructed a data set using the consensus result from NetMHCIIpan and TEPITOPEpan. The data set contains binding registers information of immunogenic and non-immunogenic HTL peptides related to HLA-DRB1*01:01. From these binding registers, we extracted various amino acid descriptor sets widely used in proteochemometric modelling (van Westen *et al*. 2013). We found that VHSE descriptor set is representative for immunogenicity modelling. A class imbalance, however, was the issue in the data set. It causes the resulted prediction model classifying unseen data as the majority class member (Kuhn & Johnson 2013). Since the model possesses a satisfactory area under ROC curve, we carried out automatic threshold probability optimization to minimize difference between specificity and sensitivity. This optimization adjusts automatically both in training set and future data. Ultimately, we have developed a random forest model of immunogenic HTL epitopes presented by HLA-DRB1*01:01. The model complements the available peptides MHC class II-binder prediction servers to generate more accurate immunogenic HTL epitopes, which further help the development of immunotherapies for HLA-DRB1*01:01-related diseases, such as multiple sclerosis, Crohn's disease, and rheumatoid arthritis.

MATERIALS AND METHODS

Data Set Preparation

Data of HLA-DRB1*01:01-related peptides complemented with T cell assays information were obtained from IEDB (http://www.iedb.org/) (Vita *et al*. 2018). We removed peptides with unnatural amino acids and duplicates. Similarly, we discarded peptides possessing both positive and negative result of T cell assays. Because of the absence of information whether peptides with negative results of T cell assays are MHC-binders, we predicted their IC50 by using NetMHCIIpan (Andreatta *et al*. 2015). Those peptides with IC50 greater than 500 nM were cut off (Karosiene *et al*. 2013; Peters *et al*. 2006) to make the data set MHC-binder exclusive. Peptide binding registers were predicted using NetMHCIIpan (Andreatta *et al*. 2015) and TEPITOPEpan (Zhang *et al*. 2012). Consensus result from both prediction servers was collected to yield a binding register data set of 392 immunogenic and 122 non-immunogenic T cell peptides related to HLA-DRB1*01:01. We then randomly split the data set into a training and test set. Amino acid descriptors of Vectors of Hydrophobicity, Steric and Electronic (VHSE) (Mei *et al*. 2005) were extracted from binding registers using a script written in R programming language (R Core Team 2015).

Random Forest

We utilized a random forest algorithm, an ensemble of decision trees (Breiman 2001), to train immunogenic HTL epitopes model. Let $\{\mathbf{x}_i\}_{i=1}^{n}$ is descriptor vectors with outcome y_i or a binary label of a peptide immunogenicity in n samples of a training data, $D = \{(\mathbf{x}_i, y_i), \ldots, (\mathbf{x}_n, y_n)\}$. The training data are drawn randomly from a probability distribution $(\mathbf{x}_i, y_i) \sim (\mathbf{X}, Y)$ to generate a random vector of descriptors Θ_k, which is independent of the previous ones $\Theta_1, \ldots, \Theta_{k-1}$. The training data and Θ_k are used to construct a classifier h(x, Θk) of the kth tree. The resulted trees vote for the most popular class at input x. Implementation of random forest in a caret package denotes the Θ_k as a tuning parameter mtry.

The Iterative Ten-Fold Cross-Validation

We performed a resampling technique, a ten-repeated ten-fold cross-validation (Kuhn & Johnson 2013), for estimating model performance. It randomly divided training set into ten subsets. The first subset was held-out, while the rest subsets fitted a model. The held-out samples were predicted by this model and used to estimate performance measures. The first subset was returned to the training set and the procedure repeated until the last subset.

Collinearity Removal

We removed collinearity within descriptors using a procedure as described in (Kuhn & Johnson 2013). The procedure, firstly, calculated correlation matrix of the predictors and determined two predictors associated with the largest absolute pairwise correlation. It then determined the average correlation between the first predictor and the other variables and did so for the second predictor. It discarded the predictor with a larger average correlation, either the first or second predictor. The algorithm iteratively run the steps above until no absolute correlations greater than a threshold of 0.750.

Recursive Feature Elimination with Resampling

We carried out recursive feature elimination, implemented in the caret package (Kuhn 2008; Kuhn *et al*. 2016), as described by Kuhn & Johnson (2013). Data were partitioned into first subset and held-back set via resampling of 10-fold cross-validation. A model was trained using all descriptors on the subset set. The hold-back samples were predicted using the model and descriptors were ranked according their importance. Another models was trained using individual descriptor and then used to predict the held-back samples. The ranking of each descriptor was recalculated. The processes were repeated for all subset. The performance profile of descriptors was calculated. The number of predictors were determined and the final list of descriptors was

estimated. The final model was then fitted based on the optimal descriptors.

Filter Method for Feature Selection

As described in (Kuhn *et al*. 2016), the algorithm used univariate statistical methods to filter descriptor variables in each iteration. It estimated the performance using resampling. In the next step, it applied the same filter and the model to entire training data. It saved the model and the current selected descriptors. The final descriptors were voted based on the optimal performance.

Threshold Probability Optimization

The optimization of threshold probability was carried out as described in (Kuhn & Johnson 2013; Kuhn *et al*. 2016). Firstly, a tuning grid searched the number of randomly selected predictors, the mtry. Using a fix mtry, the training data fitted a single random forest model. Next, the algorithm looped over the threshold values to obtain prediction from the same random forest model. It then fitted the model independent of the threshold parameter. To evaluate data across thresholds, it created multiple versions of the probabilities. Using the current candidate value of the probability threshold, it use the area under the ROC curve and the sensitivity and specificity values. At the end, it selected the threshold where the distance between sensitivity and specificity is minimum.

Averaging Probabilities

We randomly split immunogenic data in the training set into three partitions. Into each partition, we added all non-immunogenic in the training set. These steps generated three sub-groups of training set where each sub-group has different immunogenic data but same non-immunogenic ones. The sub-groups then trained different random forest learners to yield three models of HTL immunogenicity. Each model was applied on the test set and resulted in immunogenicity probabilities. We then averaged probabilities outcome from each test set data point. The final class follow the average result.

Position-Based Amino Acid VHSE Descriptors Analysis

We transformed amino acids at peptide binding registers into the VHSE descriptor set using R statistical software (R Core Team 2015). Next, we calculated and plotted the mean descriptors at each binding register residue between immunogenic and non-immunogenic peptides. In addition, we performed a Wilcoxon rank-sum test to position-based residues of between immunogenic and non-immunogenic peptides at their binding registers for each descriptor element.

Performance Metrics

Here we adopted some metrics to the model performance. They are an area under ROC curve (ROC), sensitivity or recall, specificity, positive prediction value (PPV) or precision, Mattew's correlation coefficient (MCC), Kappa, harmonic mean of precision and recall (F1), and. These metrics are defined as follows:

$$sensitivity = recall = \frac{TP}{TP + FN}$$

$$specificity = \frac{TN}{TN + FP}$$

$$PPV = precision = \frac{TP}{TP + FP}$$

$$MCC = \frac{(TP \times TN - FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$

$$Kappa = \frac{O - E}{1 - E}$$

$$F_1 = \frac{2 \times precision \times recall}{precision + recall}$$

Where TP is true positive, FN is false negative, TN is true negative, and FP is false positive. The ROC curve is obtained by plotting the false positive rate (1-specificity) against the sensitivity. The O constructing a Kappa metric is the observed accuracy, whereas E is the expected accuracy based on the marginal totals of the confusion matrix (Kuhn & Johnson 2013).

Structure and Sequence Conservation Visualization

The crystal structure of ternary complex of TCR, influenza HA antigen peptide, and HLA-DRB1*01:01 with a code 1FYT was retrieved from RCSB Protein Data Bank (Burley *et al*. 2018). The structure was visualized using Biovia Discovery Studio 2016 (Systèmes 2016). Meanwhile, sequence conservation of binding registers in data set was generated using WebLogo (Crooks *et al*. 2004).

RESULTS AND DISCUSSIONS

Screening of Descriptor Sets

Previously, Chowell *et al*. (2015) analyzed biochemical properties among immunogenic and non-immunogenic CD8+ peptides related to HLA-A2. They observed hydrophobicity differences at specific TCR contact residues P4, P6, P7, and P8. Using the hydrophobicity property, they built a neural network model to complement the IEDB approach in predicting immunogenic CTL epitopes. Khan and Ranganathan (2011) proposed that the molecular surface electrostatic potential (MSEP) contributes in pMHC recognition by TCR. Zhang *et*

al. (2015) applied various combination of physicochemical descriptors to distinguish immunogenic and non-immunogenic CTL epitopes. In the case of HTL epitopes related to HLA-DRB1*01:01, we evaluated hydrophobicity, steric, and electrostatic properties in a VHSE descriptor set (Mei *et al*. 2005) to build an immunogenic epitope prediction model. The VHSE is derived from experimental physicochemical properties of amino acids in an AAindex (Kawashima *et al*. 2008).

Feature Selection

Extraction of VHSE from binding register data resulted in 72 descriptors. These descriptors explain binding register residues in terms of hydrophobicity, steric, and electrostatic properties. During the antigen recognition event, only particular residues at the binding register interact with complementarity determining regions (CDRs) of a TCR (Sant'Angelo *et al*. 2002). Some residues at binding register might be more contribute in hydrophobicity, whereas others could participate through different physicochemical properties such as electrostatic or even do not interact with the TCR. Involving inappropriate properties, as descriptors, in predictive modelling may be redundant and decrease model performance (Kuhn & Johnson 2013; Guyon *et al*. 2006). Hence, we performed feature selection steps on the extracted descriptors to choose only the important ones.

Using a selection by filter (SBF) method, we screened important descriptors from the binding-register-extracted VHSE. The method evaluates the relevance of the predictors outside of the predictive models using univariate statistics (Kuhn & Johnson 2013). In this work, we used analysis of variance (ANOVA) score. The SBF method retains twelve descriptors (SBF1). Subjecting these descriptors to a random forest training increase the ROC to 0.701 (Table 1). The selected descriptors also increased the specificity (0.330), the precision/PPV (0.808), the Kappa (0.266), and the MCC (0.284), whereas the F1 (0.854) is unchanged.

We also carried out another feature selection, a recursive feature elimination (RFE) (Guyon *et al*. 2006). It is a backward selection algorithm that prevents refitting models at every search step. Its implementation in the caret package incorporates resampling to obtain performance estimates with variation due to feature selection (Kuhn & Johnson 2013). The use of the RFE method on the VHSE descriptors generated 15 descriptors (RFE1). Training these descriptors to a random forest algorithm increase the ROC to 0.723 (Table 1). The other increased performance metrics are the specificity (0.262), the precision/PPV (0.799), the Kappa (0.249), the F1 (0.865), and the MCC (0.299).

Separately, we applied a collinearity removal procedure to the extracted VHSE descriptors. This procedure retained 55 collinearity-free descriptors. Further SBF method selected 8 descriptors (SBF2). These descriptors increased the ROC to 0.663 (Table 1). In addition, they also increased Specificity (0.263), Precision/PPV (0.789), Kappa (0.170), F1 (0.836), and MCC (0.182).

With prior collinearity removal, recursive feature elimination retained 25 descriptors (RFE2). The use of the RFE2 increased the ROC to 0.727 (Table 1). The selected descriptors also increased Specificity (0.190), Precision/PPV (0.788), Kappa (0.209), F1 (0.870), and MCC (0.289).

The Figure 1 compares the performance of HTL epitope predictive models trained using different selected VHSE descriptors. The RFE2, VHSE descriptors selected by RFE method after collinearity removal, generated a model with the best average ROC performance (0.727). Thus, we selected the RFE2 for further modelling.

Table 1. Average performance metrics of different feature selection methods on the VHSE descriptor set. These metrics were estimated through a resampling of a ten-repeated ten-fold cross-validation during a random forest training.

	ROC	Sensitivity/ Recall	Specificity	Precision/ PPV	F1	Kappa	MCC
Full VHSE	0.672	0.954	0.144	0.775	0.854	0.124	0.176
SBF1	0.701	0.907	0.330	0.808	0.854	0.266	0.284
RFE1	0.723	0.946	0.262	0.799	0.865	0.249	0.299
SBF2	0.663	0.891	0.263	0.789	0.836	0.170	0.182
RFE2	0.727	0.973	0.190	0.788	0.870	0.209	0.289

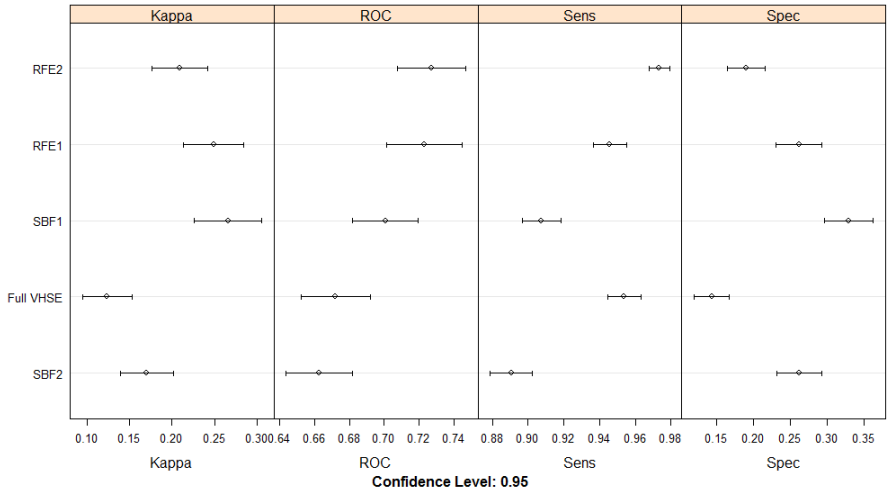


Figure 1. Average performance metrics of different feature selection methods on the VHSE descriptor set.

Table 2. Performance metrics of three sub-models and probabilities average model trained using the best selected VHSE descriptors.

Model	ROC	Sensitivity/ Recall	Specificity	Precision/ PPV	F1	Kappa	MCC
Sub-model1	0.664	0.671	0.550	0.850	0.750	0.171	0.185
Sub-model2	0.616	0.724	0.500	0.846	0.786	0.186	0.194
Sub-model3	0.692	0.658	0.600	0.862	0.746	0.194	0.214
Ensemble	0.661	0.724	0.600	0.873	0.791	0.261	0.277
Optimized probability threshold	0.672	0.750	0.600	0.877	0.809	0.291	0.304

Probability Average and Probability Threshold Optimization

The binding register data set contains an imbalance class where the immunogenic is three times of the non-immunogenic in number. To balance the data in the training set, we split the immunogenic into three different sub-groups and added all non-immunogenic data. Hence, each sub-group has different immunogenic data but same non-immunogenic ones.

We used the best selected VHSE descriptors (RFE2) in each sub-group to train three random forest sub-models. Averaging probabilities of all sub-models gave the ROC of 0.661, the sensitivity of 0.724, and the specificity of 0.600. The majority performance metrics of the average model are superior to those of the sub-models (see Table 2).

To handle the class imbalance issue, we also carried out another approach. We optimized probability threshold to get an appropriate balance between sensitivity and specificity. Such optimization tunes the model using a resampling procedure; hence no additional data set is required. Interestingly, it also automatically applies the optimized probability threshold in predicting unseen data. The threshold probability threshold procedure exhibits better performance metrics than the ensemble one, except for the specificity (Table 2). Its ROC is 0.672, whereas the sensitivity is 0.750.

Analysis of Position-Based VHSE Descriptors

To elucidate the important residue positions as well as the related physicochemical properties within the peptide binding register interacting with the TCR, we performed a Wilcoxon rank-sum test. The test result (Table 3) indicates statistical differences between the VHSE at binding registers of immunogenic and non-immunogenic. It suggests that the majority of important residues located at the center of the binding register.

The Table 3 indicates statistical differences between immunogenic and non-immunogenic peptide binding registers at positions P3, P4, P5, and P6. At these four positions, the hydrophobic principal scores (VHSE1 and 2) show statistical differences between immunogenic and non-immunogenic peptides. The immunogenic peptides have lover average VHSE1 scores at P5 (p = 2.82×10^{-2}) than the non-

Table 3. Residue-by-residue analysis of each VHSE descriptor vector between immunogenic and non-immunogenic HTL peptides at their binding registers. The analysis was determined by using the Wilcoxon rank-sum test.

	P1	P2	P3	P4	P5	P6	P7	P8	P9
VHSE1	8.10×10^{-1}	3.68×10^{-1}	5.46×10^{-1}	7.60×10^{-1}	2.82×10^{-2}	8.79×10^{-1}	7.97×10^{-1}	7.20×10^{-1}	6.70×10^{-1}
VHSE2	7.13×10^{-1}	7.58×10^{-1}	3.07×10^{-2}	8.89×10^{-3}	2.42×10^{-1}	4.89×10^{-4}	5.43×10^{-1}	7.81×10^{-1}	1.56×10^{-1}
VHSE3	4.35×10^{-1}	9.94×10^{-1}	5.52×10^{-2}	9.20×10^{-2}	3.30×10^{-1}	6.61×10^{-4}	6.46×10^{-1}	8.81×10^{-1}	5.75×10^{-1}
VHSE4	6.28×10^{-1}	6.46×10^{-1}	2.93×10^{-2}	9.58×10^{-1}	6.45×10^{-3}	7.65×10^{-2}	6.01×10^{-1}	1.16×10^{-1}	3.14×10^{-1}
VHSE5	3.09×10^{-1}	5.85×10^{-1}	1.35×10^{-1}	6.88×10^{-1}	5.58×10^{-1}	1.33×10^{-1}	7.30×10^{-1}	6.98×10^{-1}	9.69×10^{-1}
VHSE6	8.33×10^{-1}	7.00×10^{-1}	4.78×10^{-2}	5.02×10^{-1}	2.50×10^{-2}	5.83×10^{-2}	9.33×10^{-1}	7.42×10^{-1}	1.93×10^{-1}
VHSE7	6.47×10^{-1}	1.43×10^{-1}	7.77×10^{-1}	1.76×10^{-1}	2.03×10^{-1}	1.28×10^{-2}	8.94×10^{-1}	1.55×10^{-1}	4.05×10^{-1}
VHSE8	7.79×10^{-1}	5.73×10^{-1}	6.20×10^{-1}	4.90×10^{-1}	6.26×10^{-2}	4.30×10^{-2}	8.34×10^{-1}	5.39×10^{-1}	2.35×10^{-1}

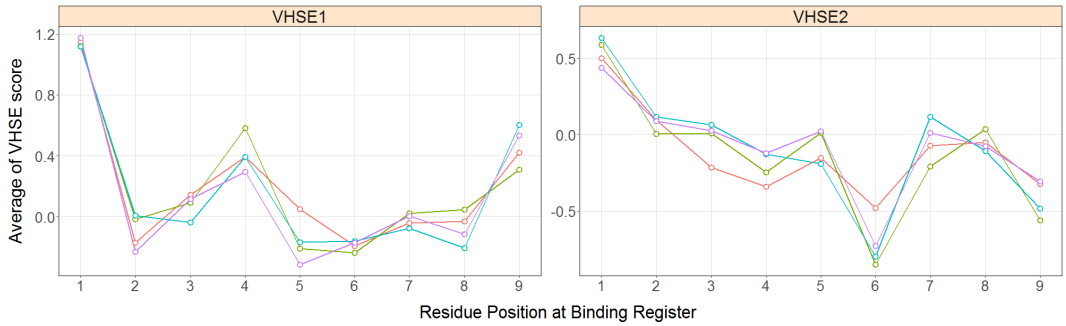


Figure 2. Comparison of average of VHSE scores representing hydrophobic properties between immunogenic and non-immunogenic at binding registers. The green, turquoise, and purple lines indicates the non-immunogenic peptides, whereas the red line is the immunogenic ones.

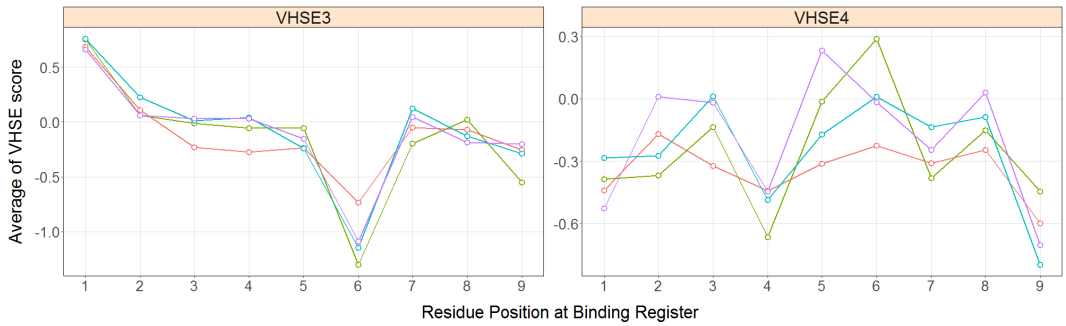


Figure 3. Comparison of average of VHSE scores representing steric properties between immunogenic and non-immunogenic at binding registers. The green, turquoise, and purple lines indicate the non-immunogenic peptides, whereas the red line is the immunogenic ones.

immunogenic ones (Figure 2). For the VHSE2, the average scores of immunogenic peptides are higher at P3 ($p = 3.07\times10^{-2}$) and P4 ($p = 8.89\times10^{-3}$), but they are lower at P6 ($p = 4.89\times10^{-4}$).

The importance of residues at middle positions is also exhibited by steric principal scores, the VHSE3 and 4. The average VHSE3 scores of immunogenic peptides are higher than that of non-immunogenic ones at P6 ($p = 6.61\times10^{-4}$) (Figure 3). In contrast, the immunogenic peptides have lower average VHSE4 scores at P3 ($p = 2.93\times10^{-2}$) and P5 ($p = 6.45\times10^{-3}$).

Of four electrostatic kind descriptors, three VHSEs (VHSE6, 7, and 8) show significant differences at middle positions of the binding register. At positions P3 and P5, average VHSE6 scores of immunogenic peptides are higher than that of non-immunogenic ones (P3, $p = 4.78\times10^{-2}$; P5, $p = 2.50\times10^{-2}$) (Figure 4). Both kinds of peptides also

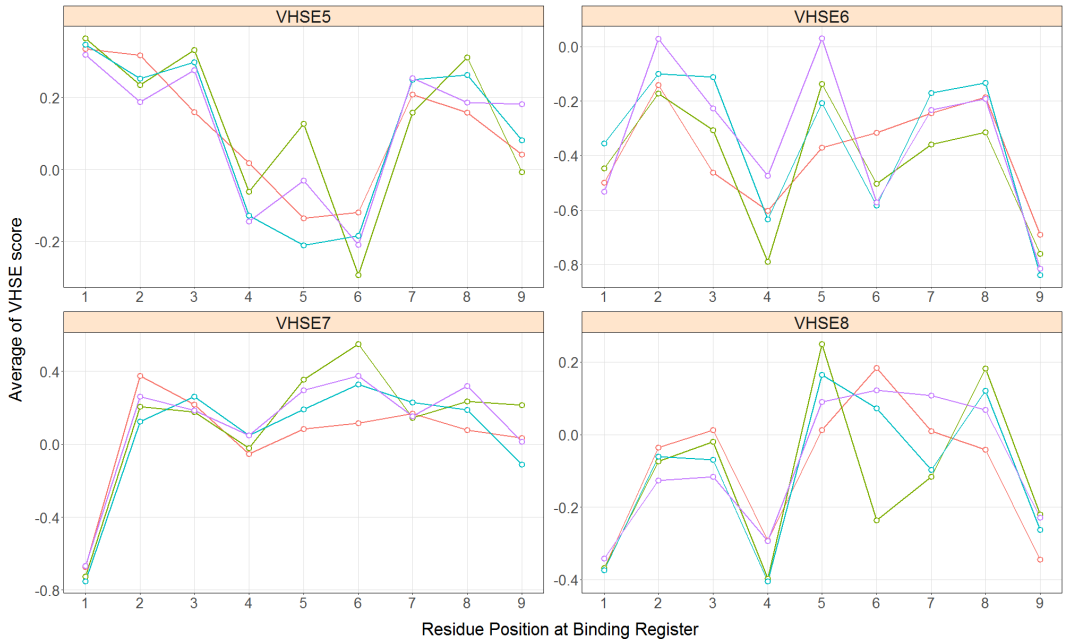


Figure 4. Comparison of average of VHSE scores representing electrostatic properties between immunogenic and non-immunogenic at binding registers. The green, turquoise, and purple lines indicate the non-immunogenic peptides, whereas the red line is the immunogenic ones.

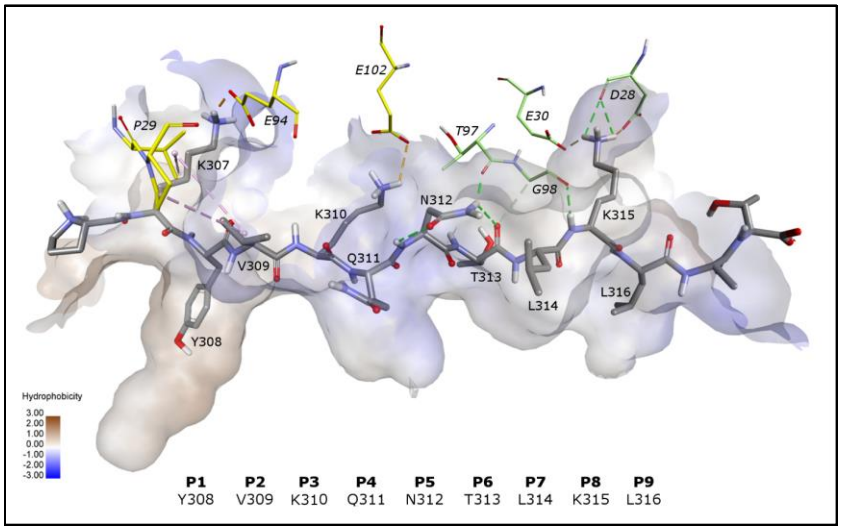
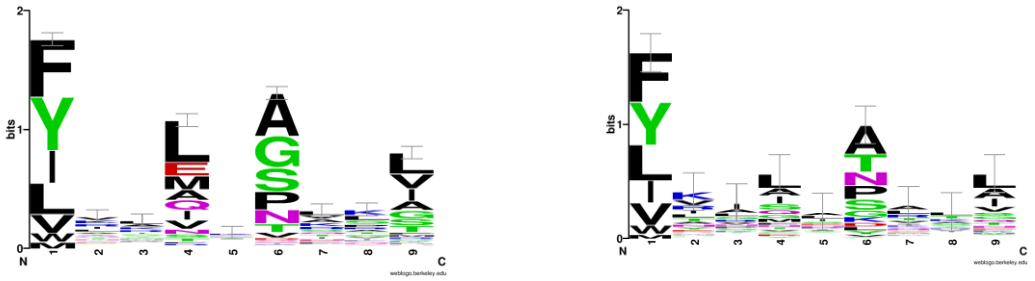


Figure 5. Interactions of a peptide epitope with residues of TCR in a crystal structure of ternary complex of TR, influenza HA antigen peptide, and HLA-DRB1*01:01 (1FYT).



Immunogenic **Non-immunogenic**

Figure 6. Sequence logos of immunogenic and non-immunogenic HTL epitopes.

have significantly different average VHSE7 and 8 scores at P6. The immunogenic VHSE7 scores are lower at this position ($p = 1.28\times10^{-2}$), whereas the VHSE8 are lower ($p = 4.30\times10^{-2}$).

The notion that important positions are at center of binding registers and the P8 is supported by a crystal structure of ternary complex of TR, influenza HA antigen peptide, and HLA-DRB1*01:01 (1FYT). This crystal structure shows that interactions occur on the peptide epitope at its center of binding register, at position P3, P5, P6, and P8. The residue at the position P3 (K310) forms a salt bridge interaction with E102 from the CDR3 of the TR, similarly the residue at P8 (K315) also interact through a salt bridge with D28 and E30 (CDR2) and a hydrogen bond with T98. However, unexpected hydrophobic interaction occurs at the P2. In addition, a hydrogen bond at flanking residue is also observed.

The Figure 6 depicts sequence logos of immunogenic and non-immunogenic HTL epitopes. Despite having high similarity at the position P1 and resemblance at other anchor residues (P4, 6, and 9), the immunogenic HTL epitopes exhibit some differences with the non-immunogenic ones at positions P4 and P6. At the position P4, an acidic residue E is the second highest occurrence in immunogenic epitopes (10.59%), whereas it only has a probability 2.41% in non-immunogenic ones. Meanwhile at the position P6, the immunogenic epitopes have higher amino acid variation (17 residues) than the non-immunogenic epitopes (12 residues). They also have high occurrence of polar residues G (20.39%) and S (15.29%), whereas in non-immunogenic are 9.64% for G and 6.02% for S. Conversely, another polar residue, T, is 15.66% in the immunogenic. It is higher than the non-immunogenic epitopes have (5.49%).

Other differences are also observed on suggested important positions for peptide-epitopes-TCR interaction. The P5 in immunogenic show high probability of charge amino acids (K and E, 11.76 and 8.63% respectively), whereas the non-immunogenic are common with hydrophobic amino acids. At the position P8, slight differences are observed. For example, immunogenic has a high probability of K 14.51%, whereas the K immunogenic is 9.64%.

CONCLUSIONS

In this work, we screened amino acid descriptor sets extracted from the peptide binding registers for immunogenicity modelling of HTL epitopes. The best performance given by VHSE suggests that hydrophobic, steric, and electrostatic properties of amino acids at the peptide binding register are sufficient for immunogenic modelling of HTL epitopes. Combination of these physicochemical properties at the center of binding register – particularly positions P3, 4, 5 and 6– and the positon P8 may play important role in the recognition of immunogenic HTL epitopes by the TCR.

In this predictive modelling of immunogenic HTL epitopes, the effect of class imbalance was persistent issue to eliminate. To alleviate this negative effect, we found that the approach of probability threshold optimization approach is better than the probabilities average.

Finally, we have developed a model for screening the immunogenic HTL epitopes from predicted peptide MHC-binders. The model helps to reduce non-immunogenic peptide MHC-binders from the result of two existing prediction webservers, NetMHCIIpan and TEPITOPEpan. Currently, the immunogenicity model is restricted to HLA-DRB1*01:01. This human MHC allele is protective against developing multiple sclerosis and association with autoimmune diseases. Thus our model is able to assist a rational development of vaccines as well as immunotherapeutic agents related to those diseases. Furthermore, the methodology can be applied to develop predictive models of immunogenic HTL epitopes restricted to another human MHC alleles.

ACKNOWLEDGEMENTS

We acknowledge the Indonesia Endowment Fund scholarship to AH and Prof. Shoba Ranganathan from Macquarie University, Australia for her supervision.

REFERENCES

Andreatta, M., Karosiene, E., Rasmussen, M., Stryhn, A., Buus, S. & Nielsen, M. (2015) Accurate pan-specific prediction of peptide-MHC class II binding affinity with improved binding core identification. *Immunogenetics*. 67(11–12): 641–650.

Bordner, A.J. & Abagyan, R. (2006) Ab initio prediction of peptide-MHC binding geometry for diverse class I MHC allotypes. *Protein: Structure, Function, and Bioinformatics*. 63: 512–526.

Breiman, L. (2001) Random forests. *Machine Learning*. 45(1): 5–32.

Burley, S.K., Berman, H.M., Bhikadiya, C., Bi, C., Chen, L., Di Costanzo, L., Christie, C., Dalenberg, K., Duarte, J.M., Dutta, S., Feng, Z., Ghosh, S., Goodsell, D.S., Green, R.K., Guranović, V., Guzenko, D., Hudson, B.P., Kalro, T., Liang, Y., Lowe, R., Namkoong, H., Peisach, E., Periskova, I., Prlić, A., Randle, C., Rose, A., Rose, P., Sala, R., Sekharan, M., Shao, C., Tan, L., Tao, Y.-P., Valasatava, Y., Voigt, M., Westbrook, J., Woo, J., Yang, H., Young, J., Zhuravleva, M., & Zardecki, C. (2018) RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Research*. 47(D1): D464–D474.

Chowell, D., Krishna, S., Becker, P.D., Cocita, C., Shu, J., Tan, X., Greenberg, P.D., Klavinskis, L.S., Blattman, J.N., & Anderson, K.S. (2015) TCR contact residue hydrophobicity is a hallmark of immunogenic CD8 + T cell epitopes. *Proceedings of the National Academy of Sciences.* 112(14): E1754–E1762.

Chuan, T.J. & Ranganathan, S. (2013) *Computer-aided vaccine design*. Woodhead Publishing.

Crooks, G., Hon, G., Chandonia, J., & Brenner, S. (2004) WebLogo: a sequence logo generator. *Genome Research*. 14: 1188–1190.

Gilboa, E. (2016) A quantum leap in cancer vaccines ? *Journal for ImmunoTherapy of Cancer*, 4: 87.

Gonzalez-Galarza, F.F., Christmas, S., Middleton, D., & Jones, A.R. (2011) Allele frequency net: A database and online repository for immune gene frequencies in worldwide populations. *Nucleic Acids Research*. 39(Supplement 1): 913–919.

Goyette, P., Boucher, G., Mallon, D., Ellinghaus, E., Jostins, L., Huang, H., Ripke, S., Gusareva, E.S., Annese, V., Hauser, S.L., Oksenberg, J.R., Thomsen, I., Leslie, S., Abraham, C., Achkar, J.-P., Ahmad, T., Amininejad, L., Ananthakrishnan, A.N., Andersen, V., Anderson, C.A, Andrews, J.M., Annese, V., Aumais, G., Baidoo, L., Baldassano, R.N., Balschun, T., Bampton, P.A, Barclay, M., Barrett, J.C., Bayless, T.M., Bethge, J., Bis, J.C., Bitton, A., Boucher, G., Brand, S., Brant, S.R., Büning, C., Chew, A., Cho, J.H., Cleynen, I., Cohain, A., Croft, A., Daly, M.J., D′Amato, M., Danese, S., De Jong, D., De Vos, M., Denapiene, G., Denson, L. a, Devaney, K.L., Dewit, O., D′Inca, R., Dubinsky, M., Duerr, R.H., Edwards, C., Ellinghaus, D., Essers, J., Ferguson, L.R., Festen, E. a, Fleshner, P., Florin, T., Franchimont, D., Franke, A., Fransen, K., Gearry, R., Georges, M., Gieger, C., Glas, J., Goyette, P., Green, T., Griffiths, A.M., Guthery, S.L., Hakonarson, H., Halfvarson, J., Hanigan, K., Haritunians, T., Hart, A., Hawkey, C., Hayward, N.K., Hedl, M., Henderson, P., Hu, X., Huang, H., Hui, K.Y., Imielinski, M., Ippoliti, A., Jonaitis, L., Jostins, L., Karlsen, T.H., Kennedy, N. a, Khan, M.A., Kiudelis, G., Kugathasan, S., Kupcinskas, L., Latiano, A., Laukens, D., Lawrance, I.C., Lee, J.C., Lees, C.W., Leja, M., Van Limbergen, J., Lionetti, P., Liu, J.Z., Louis, E., Mahy, G., Mansfield, J., Massey, D., Mathew, C.G., McGovern, D.P.B., Milgrom, R., Mitrovic, M., Montgomery, G.W., Mowat, C., Newman, W., Ng, A., Ng, S.C., Ng, S.M.E., Nikolaus, S., Ning, K., Nöthen, M., Oikonomou, I., Palmieri, O., Parkes, M., Phillips, A., Ponsioen, C.Y., Potocnik, U., Prescott, N.J., Proctor, D.D., Radford-Smith, G., Rahier, J.-F., Raychaudhuri, S., Regueiro, M., Rieder, F., Rioux, J.D., Ripke, S., Roberts, R., Russell, R.K., Sanderson, J.D., Sans, M., Satsangi, J., Schadt, E.E., Schreiber, S., Schumm, L.P., Scott, R., Seielstad, M., Sharma, Y., Silverberg, M.S., Simms, L. a, Skieceviciene, J., Spain, S.L., Steinhart, a H., Stempak, J.M., Stronati, L., Sventoraityte, J., Targan, S.R., Taylor, K.M., Velde, A. Ter, Theatre, E., Torkvist, L., Tremelling, M., van der Meulen, A., van Sommeren, S., Vasiliauskas, E., Vermeire, S., Verspaget, H.W., Walters, T., Wang, K., Wang, M.-H., Weersma, R.K., Wei, Z., Whiteman, D., Wijmenga, C., Wilson, D.C., Winkelmann, J., Xavier, R.J., Zeissig, S., Zhang, B., Zhang, C.K., Zhang, H., Zhang, W., Zhao, H., Zhao, Z.Z., Daly, M.J., Van Steen, K., Duerr, R.H., Barrett, J.C., McGovern, D.P.B., Schumm, L.P., Traherne, J.A., Carrington, M.N., Kosmoliaptsis, V., Karlsen, T.H., Franke, A., & Rioux, J.D. (2015) High-density mapping of the MHC identifies a shared role for HLA-DRB1*01:03 in inflammatory bowel diseases and heterozygous advantage in ulcerative colitis. *Nature Genetics*. 47(2): 172–179.

Guan, P., Doytchinova, I.A., Zygouri, C., & Flower, D.R. (2003) MHCPred: a server for quantitative prediction of peptide–MHC binding. *Nucleic Acids Research* . 31(13): 3621–3624.

Guyon, I., Gunn, S., Nikravesh, M., & Zadeh, L.A. (2006) *Feature Extraction: Foundations and Applications*. I. Guyon, S. Gunn, M. Nikravesh, & L. A. Zadeh, eds. Springer, Berlin.

Karosiene, E., Rasmussen, M., Blicher, T., Lund, O., Buus, S. & Nielsen, M. (2013) NetMHCIIpan-3.0, a common pan-specific MHC class II prediction method including all three human MHC class II isotypes, HLA-DR, HLA-DP and HLA-DQ. *Immunogenetics*. 65(10): 711–724.

Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., & Kanehisa, M. (2008) AAindex: Amino acid index database, progress report 2008. *Nucleic Acids Research*. 36(Supplement 1): 202–205.

Khan, J.M. & Ranganathan, S. (2010) pDOCK: A new technique for rapid and accurate docking of peptide ligands to Major Histocompatibility Complexes. *Immunome Research*. 6(Supplement 1): S2.

Khan, J.M. & Ranganathan, S. (2011) Understanding TR binding to pMHC complexes: How does a TR scan many pMHC complexes yet preferentially bind to one. *PloS One*. 6(2): e17194.

Kuhn, M. (2008) Building predictive models in R using the caret package. *Journal Of Statistical Software*. 28(5): 1–26.

Kuhn, M. & Johnson, K. (2013) *Applied Predictive Modeling*. Springer, New York.

Kuhn, M., Wing, J., Weston, S., Williams, A.,

Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., & Candan., C. (2016) Caret: Classification and regression training. Astrophysics Source Code Library.

Lerner, A., Jeremias, P. & Matthias, T. (2015) The World Incidence and Prevalence of Autoimmune Diseases is Increasing. *International Journal of Celiac Disease*. 3(4): 151–155.

Mei, H., Liao, Z.H., Zhou, Y. & Li, S.Z. (2005) A new set of amino acid descriptors and its application in peptide QSARs. *Biopolymers*. 80(6): 775–86.

Murphy, K. (2011) *Janeway's Immunobiology*. 8th ed. Garland Science. London.

Nielsen, M., Lundegaard, C., Worning, P., Hvid, C.S., Lamberth, K., Buus, S., Brunak, S. & Lund, O. (2004) Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach. *Bioinformatics*. 20(9): 1388–1397.

Oyarzún, P. & Kobe, B. (2016) Recombinant and epitope-based vaccines on the road to the market and implications for vaccine design and production. *Human Vaccines & Immunotherapeutics*. 12(3): 763–767.

Patronov, A., Dimitrov, I., Flower, D.R. & Doytchinova, I. (2011) Peptide binding prediction for the human class II MHC allele HLA-DP2: a molecular docking approach. *BMC Structural Biology*. 11(1): 32.

Peters, B., Bui, H.H., Frankild, S., Nielsen, M., Lundegaard, C., Kostem, E., Basch, D., Lamberth, K., Harndahl, M., Fleri, W., Wilson, S.S., Sidney, J., Lund, O., Buus, S. & Sette, A. (2006) A community resource benchmarking predictions of peptide binding to MHC-I molecules. *PLoS Computational Biology*. 2(6): 0574–0584.

R Core Team (2015) R: A language and environment for statistical computing. *R Foundation for Statistical Computing. Vienna, Austria*.

Rammensee, H.-G., Bachmann, J., Emmerich, N.P.N., Bachor, O.A. & Stevanović, S. (1999) SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics*. 50(3–4): 213–219.

Reche, P.A, Glutting, J.-P., Zhang, H. & Reinherz, E.L. (2004) Enhancement to the RANKPEP resource for the prediction of peptide binding to MHC molecules using profiles. *Immunogenetics*. 56(6): 405–419.

Rognan, D., Scapozza, L., Folkers, G., & Daser, A. (1994) Molecular dynamics simulation of MHC-peptide complexes as a tool for predicting potential T cell epitopes. *Biochemistry*. 33(38): 11476–11485.

Rosenfeld, R., Zheng, Q., Vajda, S. & DeLisi, C. (1995) Flexible docking of peptides to class I major-histocompatibility-complex receptors. *Genetic Analysis: Biomolecular Engineering*. 12(1): 1–21.

Saethang, T., Hirose, O., Kimkong, I., Tran, V.A., Dang, X.T., Nguyen, L.A.T., Le, T.K.T., Kubo, M., Yamada, Y. & Satou, K. (2013) PAAQD: Predicting immunogenicity of MHC class I binding peptides using amino acid pairwise contact potentials and quantum topological molecular similarity descriptors. *Journal of Immunological Methods*. 387(1–2): 293–302.

Sant'Angelo, D.B., Robinson, E., Janeway, C.A. & Denzin, L.K. (2002) Recognition of core and flanking amino acids of MHC class II-bound peptides by the T cell receptor. *European Journal of Immunology*. 32(9): 2510–2520.

Sauer, E.L., Cloake, N.C., & Greer, J.M. (2015) Taming the TCR: Antigen-specific immunotherapeutic agents for autoimmune diseases. *International Reviews of Immunology*. 34(6): 460-485.

Systèmes, D. (2016) Dassault Systèmes BIOVIA, Discovery Studio Modeling Environment, Release 2016.

Tong, J.C., Tan, T.W., & Ranganathan, S. (2007) In silico grouping of peptide/HLA class I complexes using structural interaction characteristics. *Bioinformatics (Oxford, England)*. 23(2): 177–183.

Tong, J.C., Tan, T.W., & Ranganathan, S. (2004) Modeling the structure of bound peptide ligands to major histocompatibility complex. *Protein Science*. 13(9): 2523–2532.

Tung, C.-W., Ziehm, M., Kämper, A., Kohlbacher, O. & Ho, S.-Y. (2011) POPISK: T-cell reactivity prediction using support vector machines and string kernels. *BMC Bioinformatics*. 12(1): 446.

Vita, R., Mahajan, S., Overton, J.A., Dhanda, S.K., Martini, S., Cantrell, J.R., Wheeler, D.K., Sette, A. & Peters, B. (2018) The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Research*. 47(D1), D339–D343.

van Westen, G.J., Swier, R.F., Wegner, J.K., Ijzerman, A.P., van Vlijmen, H.W. & Bender, A. (2013) Benchmarking of protein descriptor sets in proteochemometric modeling (part 1): comparative study of 13 amino acid descriptor sets. *Journal of Cheminformatics*. 5(1): 41.

WHO. 2017. *Global Health Estimates 2016 Summary*. https://www.who.int/healthinfo/global_burden_disease/GHE2016_YLD_WBI_2000_2016_.xls

Zhang, G.L., Khan, A.M., Srinivasan, K.N., August, J.T. & Brusic, V. (2005) MULTIPRED: A computational system for prediction of promiscuous HLA binding peptides. *Nucleic Acids Research*. 33(SUPPL. 2): 172–179.

Zhang, L., Chen, Y., Wong, H.-S., Zhou, S.,

Mamitsuka, H. & Zhu, S. (2012) TEPITOPEpan: extending TEPITOPE for peptide binding prediction covering over 700 HLA-DR molecules. *PloS one*. 7(2): e30483.

Zhang, W., Niu, Y., Zou, H., Luo, L., Liu, Q. & Wu, W. (2015) Accurate prediction of immunogenic T-cell epitopes from epitope sequences using the genetic algorithm-based ensemble learning. *Plos One*. 10(5): e0128194.