

EVALUATING THE PREDICTIVE PERFORMANCE OF CYTOTOXIC T LYMPHOCYTE EPITOPE PREDICTION TOOLS USING ELISPOT ASSAY DATA



By: Rebone Leboreng Meraba

Supervisor: Dr Darren Martin

Programme: Bioinformatics
Computational Biology Group
Dept of Integrative Biomedical Sciences
Faculty of Health Sciences
University of Cape Town
September 2017

The copyright of this thesis vests in the author. No quotation from it or information derived from it is to be published without full acknowledgement of the source. The thesis is to be used for private study or non-commercial research purposes only.

Published by the University of Cape Town (UCT) in terms of the non-exclusive license granted to UCT by the author.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	iv
ABSTRACT	v
LIST OF ABBREVIATIONS	vi
1. INTRODUCTION	1
1.1 Pathways that trigger an immune response	2
1.2 Structure of MHC class I molecules	3
1.3 Diversity and classification of the MHC class I molecules	5
1.4 The influence of HLA class I alleles controlling infectious diseases	6
1.5 Computational prediction of HLA class I and epitope interactions	8
1.5.1 Basic computational epitope prediction methods	9
1.5.1.1 Binding motif-based methods	9
1.5.1.2 Quantitative binding matrix-based methods	10
1.5.1.3 Machine learning-based methods	11
1.5.2 Publicly available epitope prediction tools	12
1.6 Aims and objectives	16
2. METHODS	17
2.1 System design of IMMUNO-SHARE, a web-based ELISpot data sharing resource	17
2.2 Evaluating the predictive performance of epitope prediction tools	18
2.2.1 IFN γ ELISpot assay dataset	18
2.2.2 Selecting four frequently used epitope prediction tools	18
2.2.2.1 netMHC 3.2	19
2.2.2.2 IEDB ANN	19
2.2.2.3 IEDB ARB Matrix	20
2.2.2.4 IEDB SMM	20
2.2.3 Compiling the testing datasets	21
2.2.4 Evaluating the predictive performance of the tools using the testing dataset	22
2.2.5 Filtering each tool's prediction results	22
2.2.6 Regression analysis of the SFUs and the quantitative predictions of the tools	23
2.2.7 Customized computer scripts	24

3. RESULTS	25
3.1 IMMUNO-SHARE, a web-based ELISpot data sharing resource	25
3.1.1 System components and features	26
3.1.1.1 Uploading procedure	26
3.1.1.1.1 File uploads	26
3.1.1.1.2 Additional data entry	28
3.1.1.1.3 Uploading options	28
3.1.2 Automatically generated files	30
3.1.3 Downloading files	31
3.2 Testing the predictive power of epitope detection tools in the context of ELISpot experiments	31
3.3 Testing the correlation between the SFUs and the strongest predicted binding affinity scores across the four prediction methods	34
4. DISCUSSION AND CONCLUSION	36
5. FUTURE WORK	42
REFERENCES	43
APPENDICES	
• APPENDIX A	
1. A0101	
2. A0201	
3. A0301	
4. A2402	
5. A2601	
6. B44	
7. B0801	
8. B1501	
9. B2705	
10. B3901	

- APPENDIX B
 - 1. netMHC_3.2_Output
 - 2. IEDB_ARB_Matrix_Output
 - 3. IEDB_ANN_Output
 - 4. IEDB_SMM_Output
- APPENDIX C
 - Duplicates
 - Example_HLA_Information
 - Example_Layout_Duplicates
 - Example_SFU_Data
 - No_Replicates
 - Example_HLA_Information
 - Example_Layout_Duplicates
 - Example_SFU_Data
 - Triplicates
 - Example_HLA_Information
 - Example_Layout_Duplicates
 - Example_SFU_Data

ACKNOWLEDGEMENTS

My sincere appreciation for the blessings bestowed upon me through the unwavering patience,
support and guidance of my supervisor.

I owe my deepest gratitude to the one who has always believed in me. To my supreme King, I thank
you for your faithfulness and merciful flowing streams of strength and comfort, which has enabled
me to soldier on during the best of times and the worst of times.

To my sources of inspiration: the Meraba family - Thomas, Jacobeth and Gaopalelwe; the Makokwe
family - Hophney, Mmaphuthi, Onalerona and Omolemo; the Mariaye family – Sandy; and the
Moshe family – Gaolekwe and Galaletsang, I am truly humbled by your infinite motivation and
encouragement.

Lastly, to my knowledgeable friends - Mariba Lebeko, Brejnev Muhire and Kamogelo Lebeko, I am
grateful for all of your efforts and assistance.

It is done and for that, I thank you.

ABSTRACT

Computational T-cell epitope prediction tools have been previously devised to predict potential human leukocyte antigen (HLA) binding peptides from protein sequences. These tools are complements of Enzyme-linked immunosorbent spot (ELISpot) assays – a very commonly applied immunological technique that is used both to identify regions of pathogen genomes that trigger an immune response and to characterize the relationships between an individual's complement of HLA alleles and the degree of immunity that they display. If computational tools could accurately predict HLA-peptide binding, then these tools might be useable as a cheap and reliable alternative to ELISpot assays.

A web-based IFN γ ELISpot assay dataset sharing resource, called IMMUNO-SHARE, was developed to enable the simple and straightforward storage and dissemination amongst researchers of large volumes of IFN γ ELISpot assay data. Such experimental data was next used to make HLA-peptide binding predictions with four frequently used T-cell epitope prediction tools – netMHC 3.2, IEDB_ANN, IEDB_ARB Matrix and IEDB_SMM. The predictive performances of all four tools individually and collectively was statistically assessed using non-parametric Spearman rank-order correlation tests.

It was found that none of the four tested tools yielded binding affinity predictions that were detectably correlated with the observed ELISpot data. High false positive rates, where high predicted binding affinities between peptides and patient HLAs corresponded in these patients with no appreciable immune responses, were apparent for all four of the tested methods.

The low degree of correlation between ELISpot data and HLA-peptide binding predictions and in particular, high false positive rates and relatively low true positive and true negative rates, indicate that the four tested tools would require substantial improvement before they could be seen as a viable alternative to ELISpot assays. Given that the accuracy of predictions of each of the four methods tested is largely dependent on both the quantity and quality of known true binder and true non-binder datasets that were used to train the HLA-peptide binding prediction methods implemented by the tools, it is plausible that the accuracy of these tools could be increased with larger training datasets. Retraining either the current methods or the next generation of prediction tools would therefore be greatly facilitated by the availability of large quantities of publically available HLA-peptide binding interaction information. It is hoped that IMMUNO-SHARE or some other ELISpot data sharing resource could eventually meet this need.

LIST OF ABBREVIATIONS

Abbreviation	Definition
MHC	Major Histocompatibility Complex
HLA	Human Leukocyte Antigen
CTL	Cytotoxic T Lymphocytes
HIV-1	Human Immunodeficiency Virus type 1
ER	Endoplasmic reticulum
TAP	Transporter associated with antigen processing
WHO	World Health Organization
HCV	Hepatitis C virus
ARV	Antiretroviral
PBMC	Peripheral blood mononuclear cells
ELISpot	Enzyme-linked immunosorbent spot
ELISA	Enzyme-linked immunosorbent assay
SFU	Spot forming unit
SFC	Spot forming cell
ROC	Receiver Operating Characteristic
SE	Sensitivity
SP	Specificity
TP	True Positive
FN	False Negative
FP	False Positive
TN	True Negative
ANN	Artificial Neural Networks
ARB	Average Relative Binding
SMM	Stabilized Matrix Method
HMM	Hidden Markov Models
SVM	Support Vector Machine
IEDB-AR	Immune Epitope Database and Analysis Resource
QM	Quantitative Matrices
HTML5	HyperText Markup Language 5
CSS	Cascading style sheets
OLP	Overlapping long peptides
nM	nanomolar units
SB	Strong binders/epitopes
WB	Weak binders
NB	Non-binders/non-epitopes
BLOSUM62	Block Substitution Matrix 62
PC	Positive control
NC	Negative control
.xls	excel file format/extension
.txt	text file format/extension
.png	portable network graphics

1. Introduction

Major Histocompatibility Complex (MHC) molecules are proteins that are expressed on the surface of leukocytes (white blood cells). In humans, the MHC encoding genes are located on the short arm of chromosome 6 at a location known as the Human Leukocyte Antigen (HLA) region. Molecules within this HLA region mediate cytotoxic immune responses [1]. These responses depend on the ability of the host's antigen-specific cells, Cytotoxic T Lymphocytes (CTL) - also called CD8⁺ T cells - to recognize pathogen-derived proteins which are presented to them by HLA molecules (Figure 1.1). Subsequent to their recognizing pathogen-derived proteins, CD8⁺ T cells attempt to prevent further spread of pathogenic microbes within the host by releasing cytotoxic chemicals (called cytokines) which kill infected cells.

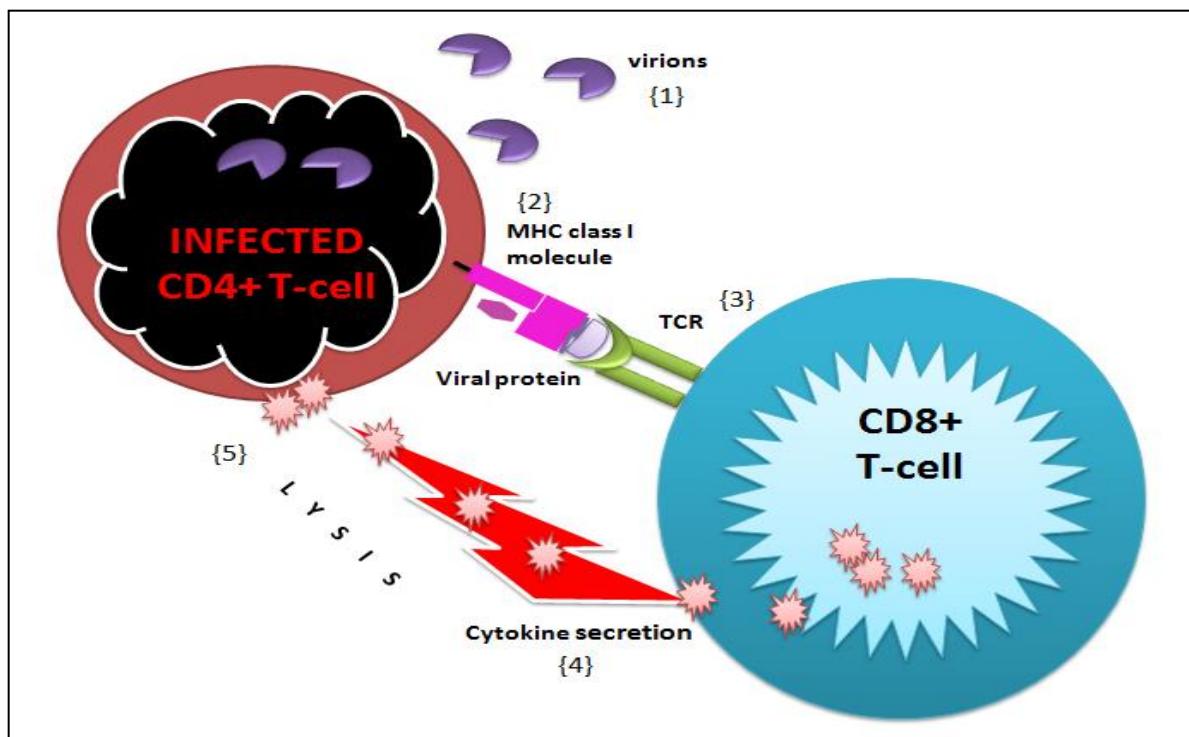


Figure 1 .1: Schematic diagram of an infected cell triggering an immune response. (1) The viral peptides infect the CD4⁺ T-cells. (2) The viral peptide is presented by a HLA class I molecule on the surface of the infected cell. (3) Recognition of the viral peptide by the CD8⁺ T-cell is mediated by the T cell receptor (TCR). (4) This recognition triggers the CD8⁺ T-cell to release cytotoxic chemicals (cytokines) which (5) kill the infected cell (Figure from [1]).

1.1 Pathways that trigger an immune response

The precise execution of two major pathways - antigen processing and peptide presentation - is required to enable CD8⁺ T-cells to recognize pathogen infected cells (Figure 1.2).

The fundamental purpose of the antigen processing pathway is to generate epitopes – the peptides that ultimately trigger an immune response. The generation of epitopes, involves two major steps: proteasomal cleavage and peptide transport [2-4]. Within a human immunodeficiency virus type 1 (HIV-1) infected cell, for example, proteasomal cleavage occurs subsequent to the virus producing its own proteins (Gag, Pol, Env, Tat, Vpr, Vif, Rev, Nef, Vpu). These proteins get cleaved into smaller fragments (i.e. peptides) by host protein degrading complexes (called proteasomes). This process yields subsets of peptides from eight to eleven amino acids in length (also called 8 to 11mers). These peptides are then transported from the cytoplasm into the endoplasmic reticulum (ER) by the transporter associated with antigen processing (TAP) molecule, leading to peptide presentation: at which point binding between the peptides and MHC-I molecules can occur [5].

The ability of CD8⁺ T-cells to recognize a specific peptide relies on the affinity with which HLA molecules bind the peptide. It has been experimentally demonstrated that in order for a CTL response to be elicited, the binding affinity must be within a specific threshold range [6, 7].

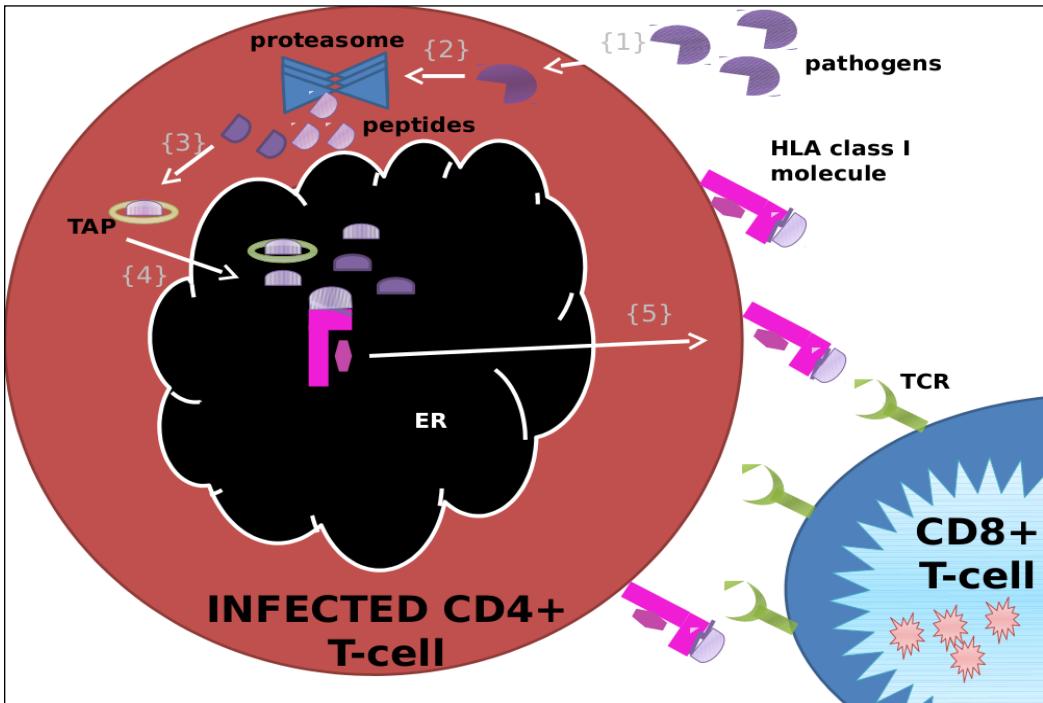


Figure 1.2: Schematic diagram illustrating the antigen processing and peptide presenting pathways. Antigen processing pathway: (1) Subsequent to a pathogen infection, (2) the host's proteasome cleaves the pathogen's proteins into peptides. (3) These peptides are then transported to the endoplasmic reticulum (ER) by the transporter associated with antigen processing (TAP) molecule. (4) Execution of this peptide transport process ensures binding between the peptides and the HLA class I molecule. The binding affinity between the peptide and HLA class I molecule can either be high (light purple) or low (dark purple). Peptide presenting pathway: the type of binding affinity is important because it affects the ability of the CD8⁺ T-cells to recognize (5) the peptide presented by the HLA class I molecule (Figure from [1]).

1.2 Structure of MHC class I molecules

MHC encoding genes are categorized within one of the three MHC classes: MHC class I, MHC class II and MHC class III. The main focus of this study is on MHC class I (MHC-I) molecules, since immunity is triggered by these molecules presenting peptides to CTLs. Whereas MHC class II molecules present peptides to helper T-cells (a type of leukocyte) which activates B-cells (another type of leukocyte) so as to induce antibody secretion, MHC class III molecules regulate inflammation and other immune system processes [8].

MHC class I molecules are composed of three heavy (α) chain domains (called α_1 , α_2 and α_3) and a single light chain (called $\beta 2$ -microglobulin) which is encoded on chromosome 15 (Figure 1.3) [1]. The top surface between the α_1 and α_2 domains forms a furrow known as the binding groove. This binding groove is the site at which viral proteins bind to the MHC-I molecule to enable viral protein presentation to the $CD8^+$ T-cells.

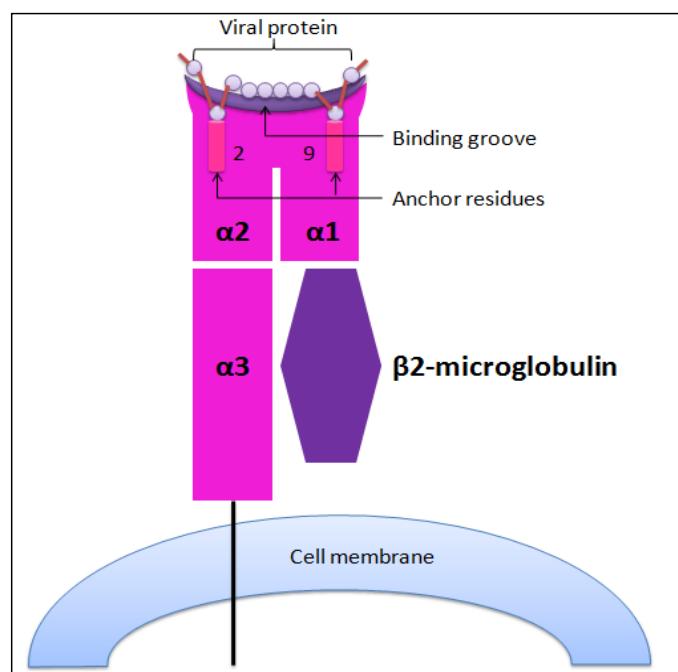


Figure 1.3: A diagram indicating the structure of a MHC class I molecule which consists of three heavy (α) chain domains (called α_1 , α_2 and α_3) and a single light chain (called $\beta 2$ -microglobulin) which is encoded on chromosome 15. Between α_1 and α_2 , is the binding groove which enables the pathogen-derived protein to bind to the MHC class I molecule (Figure from [1]).

1.3 Diversity and classification of the MHC class I molecules

There are three main types of MHC-I molecules, called HLA-A, HLA-B and HLA-C, all of which are encoded by different genes. In addition to these, there are 14 pseudogenes and three genes encoding “non-classical” MHC-I molecules, called HLA-E, HLA-F and HLA-G [1].

HLA-A, -B and -C (i.e. the “classical” HLA molecules) are essential in immune responses because they present antigenic peptides to CTLs, whereas the role of the non-classical HLA molecules is more restricted [9]. Due to the essential role of the classical molecules in immune responses, they have been far more extensively studied than the non-classical ones. These studies have revealed a vast degree of diversity in the amino acid sequences of the class I α chain such that the genes encoding MHC molecules are amongst the most polymorphic regions in the human genome.

According to the HLA nomenclature (<http://hla.alleles.org/>) there are 10 163 identified classical MHC-I gene alleles. The most polymorphic of these classical genes, HLA-B, has 4 077 known alleles whereas there are 3 285 HLA-A alleles and 2 801 HLA-C alleles [10]. Due to their large numbers, the World Health Organization (WHO) Nomenclature Committee for Factors of the HLA System has specifically formulated a naming system to differentiate between all of these HLA alleles [11, 12]. To illustrate, an HLA-B allele, HLA-B*15:01, is identified by the asterisk (*) and the two sets of unique numbers separated by the colon provides detailed information on the HLA allele. In this case, set one indicates the allele group, 15, and set two indicates the specific HLA allele, 01.

In addition, HLA alleles are clustered within groups referred to as supertypes. These supertypes indicate the HLA alleles which share specific antigenic binding properties [13]. Such properties depend on the physicochemical function of the type of amino acid at particular so-called “anchor residues” within an antigenic sequence. Anchor residues are specific amino acid positions (mostly positions two

and nine) that firmly secure the antigen within the HLA binding groove and have a disproportionately large influence on the ability of an antigen to bind to a specific HLA molecule [14].

There are also various HLA supertypes such as A1, A2, A24, B7, B44 and B58 [15]. For instance, the A1-supertype includes several HLA-A alleles (HLA-A*01:01, -A*26:01, -A*29:02 and -A*30:02) which bind peptides containing small or aliphatic amino acids in position two and aromatic amino acids at the C terminus [16]. The B44-supertype on the other hand consists of a collection of HLA-B alleles (HLA-B*1801, -B*40:01, -B*40:02, -B*44:03 and -B*45:01) which recognize peptides that have acidic (glutamic and aspartic acid) amino acids in position two and hydrophobic or aromatic amino acids at the C terminus [17].

As a result of the extreme diversity of the classical MHC class I genes, each human carries their own rare (and possibly unique) combination of HLA alleles [18]. Although each individual inherits a pair of each of the three classical MHC class I genes, the combinations of HLA alleles can vary widely even between closely related individuals [1].

1.4 The influence of HLA class I alleles in controlling infectious diseases

Susceptibility and resistance to numerous infectious diseases such as HIV-1 and hepatitis C virus (HCV) have been associated with the polymorphic nature of the HLA complex.

For example, various studies have indicated that certain HLA alleles are significantly associated with either slower or faster HIV-1 disease progression [19-24]. Specifically, whereas HLA-A*24, HLA-B*8, HLA-B*35, HLA-B*37, HLA-B*53:01, HLA-B*56, HLA-C*04 are associated with more rapid disease progression, HLA-A*2, HLA-B*27, HLA-B*51, HLA-B*53, HLA-B*57:01, HLA-B*58:01, HLA-C*08 and HLA-C*14 are associated with slower disease progression.

Similarly, it has been observed that HLA-A*2301, HLA-B*18 and HLA-C*04 are associated with HCV disease progression, whilst resistance to HCV is associated with HLA-A*03, HLA-B*27, HLA-B*57 and HLA-C*01 [25].

Ultimately, identifying how a HLA molecule is associated with the progression of an infectious disease enables the identification of regions within a pathogen's genome which bind to these HLA molecules. For instance, one study identified the epitopes within the HIV-1 subtype B proteome that are predominately presented by HLA molecules which are associated with either slower or faster disease progression [26]. Whereas HLA molecules associated with slower disease progression bound to Gag-p24 peptides, those associated with a faster disease progression predominantly bound Nef peptides. Furthermore, another study [27] examined the difference in immunological control between (1) slow progressors (i.e. HIV-1 infected individuals who manage to control the infection for 10 years or more), (2) antiretroviral (ARV) treated progressors and (3) ARV untreated progressors, all of which carried the HLA-B*57:01 allele. Each patient's peripheral blood mononuclear cells (PBMC) were tested for the presence of three particular HIV-1 Gag-p24 epitopes which are known to bind to HLA-B*57:01: ISPRTLNAW (IW9, Gag₁₄₇₋₁₅₅), KAFSPEVIPMF (KF11, Gag₁₆₂₋₁₇₂) and QASQEVKNW (QW9, Gag₃₀₈₋₃₁₆). It was discovered that whereas all these epitopes within the slow progressor group were being presented, only the KF11 and IW9 epitopes were being presented within the untreated progressor group and only the KF11 epitope was being presented within the treated progressor group.

In conclusion, being able to identify the HLA molecule and epitope combinations which mediate the best immune responses against an infectious disease is essential as such knowledge will help inform the development of CTL-epitope-based vaccines [28].

1.5 Computational prediction of HLA class I and epitope interactions

Computational epitope prediction tools have the potential to complement immunological techniques for identifying epitopes such as Enzyme-Linked ImmunoSpot (ELISpot) [29] and Enzyme-linked immunosorbent assay (ELISA) [30]. These laboratory techniques are primarily designed to monitor immune responses and identify both regions of pathogen genomes that trigger immune responses, and the relationships between HLA alleles and immunity [1]. Since these laboratory techniques are expensive and time consuming, accurate computational methods for predicting HLA-epitope interactions and binding affinities would be extremely valuable.

Given that data generated from laboratory-based investigations of HLA binding have in the past had a major impact on epitope-based vaccine development [31, 32], so too could accurate computational tools for predicting HLA-epitope interactions. Ideally, faster and cheaper identification of CTL epitopes that are suitable for inclusion within vaccines will have a positive effect on the rate at which effective individualised vaccines could be designed and manufactured.

To fill this need, a range of epitope prediction tools have been developed. With the aid of these tools, identifying T-cell epitopes has become more efficient by minimizing the numbers of *in vitro* tests that need to be carried to identify peptides that are likely to bind to specific HLA alleles [1]. It is important to note, however, that although there is widespread acceptance of the value of computational epitope predictions, in the absence of solid experimentation, such predictions must always remain hypothetical. It is because of this that the utility of these tools remains restricted to focusing laboratory-based binding experiments on particularly plausible epitope-HLA combinations [1].

1.5.1 Basic computational epitope prediction methods

In the following sections, I will describe the various prediction methods employed by the different epitope prediction tools. There are two main categories of prediction methods: pattern identifying methods and 3D structure oriented methods [33 - 41]. Whereas techniques in the latter category can be extremely accurate, they are complex and computationally expensive and, possibly because of this, no widely applicable tools have yet been developed. Most epitope prediction tool development has focused on less computationally expensive methods which seek to identify patterns of amino acids within epitopes that are indicative of an affinity for particular HLA alleles. There are three basic categories of these pattern identifying methods: (1) binding motif-based, (2) quantitative binding matrix-based and (3) machine learning-based [1].

1.5.1.1 Binding motif-based methods

These methods rely on the accurate identification of anchor residues within peptide sequences. The frequency of each amino acid at any given peptide position is represented in the form of a binding matrix: an experimentally informed data construct (Figure 1.4) [1]. Each HLA allele has its own binding matrix which is used to predict binding between a peptide and that respective allele. When predicting binding between a peptide and an allele, binding motifs within the peptide are searched for, and the success or failure of detecting the expected amino acids at the expected anchor residue locations (which are referenced from the allele's binding matrix) indicates the peptide's binding ability [1].

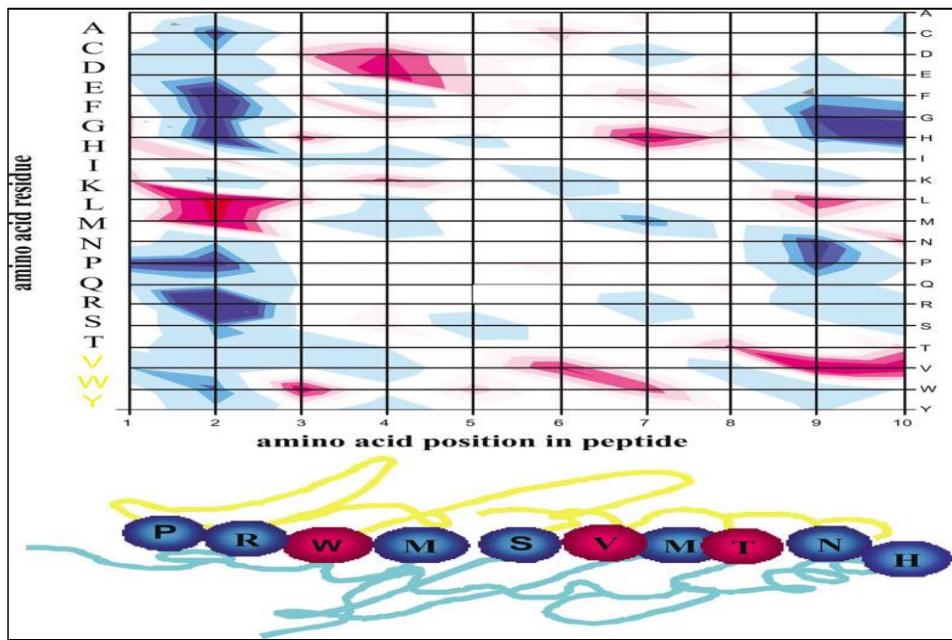


Figure 1.4: An illustration of the binding matrix-based method. The top figure is the MHC class I binding matrix for HLA A*0201. This matrix indicates the frequency of all the amino acids (Y-axis) at a given position in the peptide sequence (X-axis). The frequency levels are represented by different colours; an amino acid that is frequently found at a given peptide position is represented in magenta, while a less frequent amino acid is represented in blue. The bottom figure is a classic ten amino acid peptide sequence, with each amino acid being coloured according to their frequency level as illustrated in the binding matrix. (Figure from [36])

1.5.1.2 Quantitative binding matrix-based methods

Quantitative binding matrix-based methods are extensions of binding motif-based methods. Rather than making binary qualitative predictions as to whether a particular peptide will bind to a given HLA allele, these methods attempt to quantify the affinity of such bindings so as to improve the accuracy with which epitopes can be predicted [1]. Specifically, these methods calculate a score for each amino acid at any given position in the peptide sequence. Each score is then represented in the form of a matrix which is used to make a quantitative prediction of peptide binding [42].

1.5.1.3 Machine learning-based methods

In supervised machine learning, training a method involves using experimental data (i.e. a training dataset) to prepare the method so that it will perform accurately on new data (i.e. test datasets). The accuracy with which associations that are detected with training datasets can be generalised to identify HLA-epitope interactions in the test datasets is dependent on the training process [1].

Artificial Neural Networks (ANNs), Hidden Markov Models (HMMs) and Support Vector Machine (SVM) approaches are some the better known machine learning-based methods which have been used so as to improve the accuracy of T-cell epitope predictions [43 - 46].

As these methods owe their success to the use of experimental training data, it is important to consider that raw experimental data may present certain complexities, such as data quality and quantity, which must be suitably addressed to ensure the success of these methods. Failure to filter out biased data and account for data limitations can compromise the utility of these methods. Biased training data can result in these methods becoming “over-trained” to the point that they will confidently infer an absence of binding between strongly binding HLA-epitope pairs that were not closely related to those in the training dataset. Also, the data used in training will always be imperfect such that data errors, or even unavoidable statistical noise, will undermine the accuracy of the methods that it is used to train. Finally, since training datasets must by necessity be derived directly from experimental data, the biggest technical issue with assembling a training dataset is performing enough HLA-peptide binding assays to produce sufficiently detailed datasets. Given that the most important limitation on the predictive power of these methods is the amount of data that is available to train them, the production of large publically available datasets detailing the binding-affinities (or even just the binary binding/non-binding status) of a wide variety of HLA-epitope pairs would be a major advance.

1.5.2 Publicly available epitope prediction tools

There are currently more than 25 publicly available T-cell epitope prediction tools which employ a variety of different epitope prediction methods. Although certain tools employ the same or similar methods, they differ with respect to: (1) the number of HLA-alleles for which epitopes can be predicted, (2) the lengths of peptides that they are capable of testing, and (3) the numbers of peptide-HLA pairs used for training the methods (Table 1.1) [1]. Ideally, a prediction tool should allow testing for a variety of HLAs with any peptide sequence that falls within the range of known epitope lengths (between 8 and 12 amino acids).

With the availability of so many epitope prediction tools, users are challenged with carefully selecting the appropriate tool for their research. Consequently, studies which evaluate and compare the predictive performances of these tools seek to inform the choices that users must make. For example, one study [76] used independent experimentally derived test data to evaluate the accuracy of 30 T-cell epitope prediction tools. Among the best performing tools based on receiver operating characteristic (ROC) analysis (which measures accuracy based on sensitivity [SE] and specificity [SP] statistics [77, 78]; refer to Table 1.2) were the tools, netMHC_ANN and IEDB_ANN, as they had reasonably balanced SE and SP scores. In contrast, the tools IEDB_SMM and IEDB_ARB yielded high SE scores but frequently identified non-binders as binders (i.e. they had high false positive rates and therefore yielded low SP scores).

In another study [79] that used test data sourced from a community binding resource and the literature, netMHC 3.0 ANN, IEDB_SMM and IEDB_ARB were amongst the 16 T-cell epitope prediction tools evaluated for their ability to discriminate between binders and non-binders of HLA-A*02:01. In this analysis, a comparison of the predictive performance – based on SE scores – ranked netMHC_ANN as the best tool for identifying actual binders, followed by IEDB_SMM and IEDB_ARB. Furthermore, a

comparison of each tool's SP scores ranked netMHC_ANN as the best tool for identifying actual non-binders, followed again by IEDB_ARB Matrix and IEDB_SMM.

To summarize, due to possible overlaps between the training and testing datasets or potential biases in each tool's training dataset, conclusions about the accuracy of the tools that were evaluated in these studies remain tentative. It is for this reason, that in my study I will use an alternative ELISpot-based approach to evaluate the performance of four of the most frequently used (and freely available) epitope prediction tools: netMHC 3.2, IEDB_ANN, IEDB_ARB and IEDB_SMM.

Table 1.1: List of several publically available T-cell epitope prediction tools including the prediction method(s) employed; the number of HLA alleles for which epitopes can be predicted and the lengths of peptides that they are capable of testing.

No.	Tool Name	Method	HLA Class I Alleles*	Peptide Length	URL Address	Reference
1	NetMHC 3.2 ^A	Artificial Neural Networks (ANNs)	78	8-14mer	http://www.cbs.dtu.dk/services/NetMHC	[45, 47-49]
		1. ANN ^B	47	8-14mer		[45]
2	IEDB	2. ARBmatrix: Quantitative matrix ^C	56	9,10mer	http://www.iedb.org/	[50]
		3. Stabilized Matrix Method (SMM) ^D	48	8-11mer		[51]
		4. NetHMPpan	71	8-14mer		[52, 53]
		5. IEDB Recommended: Uses the consensus, ANN, SMM, NetMHCpan, Comblib methods	71	8-14mer		[54]
		6. Consensus: Combines: ANN, SMM, Comblib_Sidney2008	48	8-14mer		
		7. Comblib_Sidney2008	12	9mer		[55]
		8. SMMPMBEC	48	8-11mer		[56]
3	BIMAS	Quantitative Matrices (QM)	33	8-10mer	http://www-bimas.cit.nih.gov/molbio/hla_bind/	[57, 58]
4	MMBPred	QM	39	9mer frame	www.imtech.res.in/raghava/mmbpred/	[59]
5	NetCTL 1.2	Weight matrices	12 Supertypes	9mer frame	http://www.cbs.dtu.dk/services/NetCTL	[45, 60 - 64]
6	nHLAPred	ANN	26	9mer frame	http://www.imtech.res.in/raghava/nhlapred/	[65]
		ComPred: QM and ANN	58			
7	ProPred-1	QM	40	9mer frame	http://www.imtech.res.in/raghava/propred1	[66]
8	RankPep	QM	75	8-11mer	http://immunax.dfci.harvard.edu/Tools/rank_pep.html	[67]
9	SVMHC	Support Vector Machine (SVMHC)	26 MHCBN-based data 13 SYFPEITHI-based data	10-9mer 8-10mer	http://www.sbc.su.se/~pierre/svmhc/	[68 , 69]
10	SYFPEITHI 1.0	Binding Motif-based Matrices	33	8-11mer	http://www.syfpeithi.de	[70]
11	WAPP	SVMHC	4	9mer	http://abi.inf.uni-tuebingen.de/Services/WAPP/	[71]
12	HLAPred	QM	54	10-12mer	http://www.imtech.res.in/raghava/hlapred/	[72]
13	Tmhcpred	QM	39	10-12mer	http://www.imtech.res.in/raghava/tmhcpred/	[73]
14	MAPPP	QM (BIMAS) Binding Motif-based matrices (SYFPEITHI1.0)	35	8-10mer	http://www.mpiib-berlin.mpg.de/MAPPP/binding.html	[74]
15	NetTepi 1.0	An integrated method: combine 3 prediction methods: 1.pMHC binding affinity, 2. pMHC stability, 3. T cell propensity	13	8-14mer	http://www.cbs.dtu.dk/services/NetTepi/	[75]

* Number of MHC class I data for human species only.

^{A-D} The four prediction tools being analysed in this study.

Table 1.2: Statistical descriptions of computational predictions with respect to each experimental dataset (Figure from [1])

		Experimental Dataset	
		True Positive List	True Negative List
Computational Predictions	Positive Predictions	TP	FP
	Negative Predictions	FN	TN
Total		TP + FN	FP + TN
ROC analysis	Sensitivity	$SE = TP / (TP + FN)$	
	Specificity		$SP = TN / (FP + TN)$

TP: True Positive; FN: False Negative; FP: False Positive; TN: True Negative.

Specifically, this study will explore: (1) the design and development of a web-driven ELISpot data sharing resource and (2) the use of ELISpot data to test and evaluate the predictive performance of these four tools. The latter involves several consecutive steps: (a) the sourcing of ELISpot data; (b) the selection of the four MHC class I CTL epitope prediction tools; (c) the installation of four tools on a local network cluster; (d) the evaluation of the predictive performances of the CTL prediction tools using ELISpot data.

1.6 Aims and objectives

The project sets out to (1) design and develop IMMUNO-SHARE, an efficient and easily accessible web-based ELISpot data sharing resource and (2) use ELISpot assay data to evaluate the predictive performance of four epitope prediction tools

Objectives:

1. Design and develop IMMUNO-SHARE, a web-based ELISpot data sharing resource.

Manage the storing and sharing of ELISpot assay data in order to ensure that users are able to easily upload, browse and download ELISpot data.

2. Assess the predictive capability of four frequently used epitope prediction tools.

Use ELISpot assay data to test and evaluate the ability of these tools to accurately identify actual binders and non-binders.

My hypothesis is that each tool's quantitative and qualitative predictions will, to a degree that is indicative of its accuracy, correlate with the unprocessed spot forming unit (SFU) counts that are yielded by IFN γ ELISpot assays.

2. METHODS

2.1 System design of IMMUNO-SHARE, a web-based ELISpot data sharing resource.

The design and presentation of IMMUNO-SHARE was developed using a combination of HyperText Markup Language 5 (HTML5), Cascading Style Sheets (CSS), JavaScript 1.5 (<https://www.javascript.com>), Ajax and JQuery-1.10.1 (<https://jquery.com>).

Django v1.7, a web framework which is written in python (<https://www.djangoproject.com>), was used to develop IMMUNO-SHARE's system functionality. In addition, the relational database (as illustrated in Figure 2.1) which is used to store and access data was implemented with PostgreSQL 9.1 (<https://www.postgresql.org>).

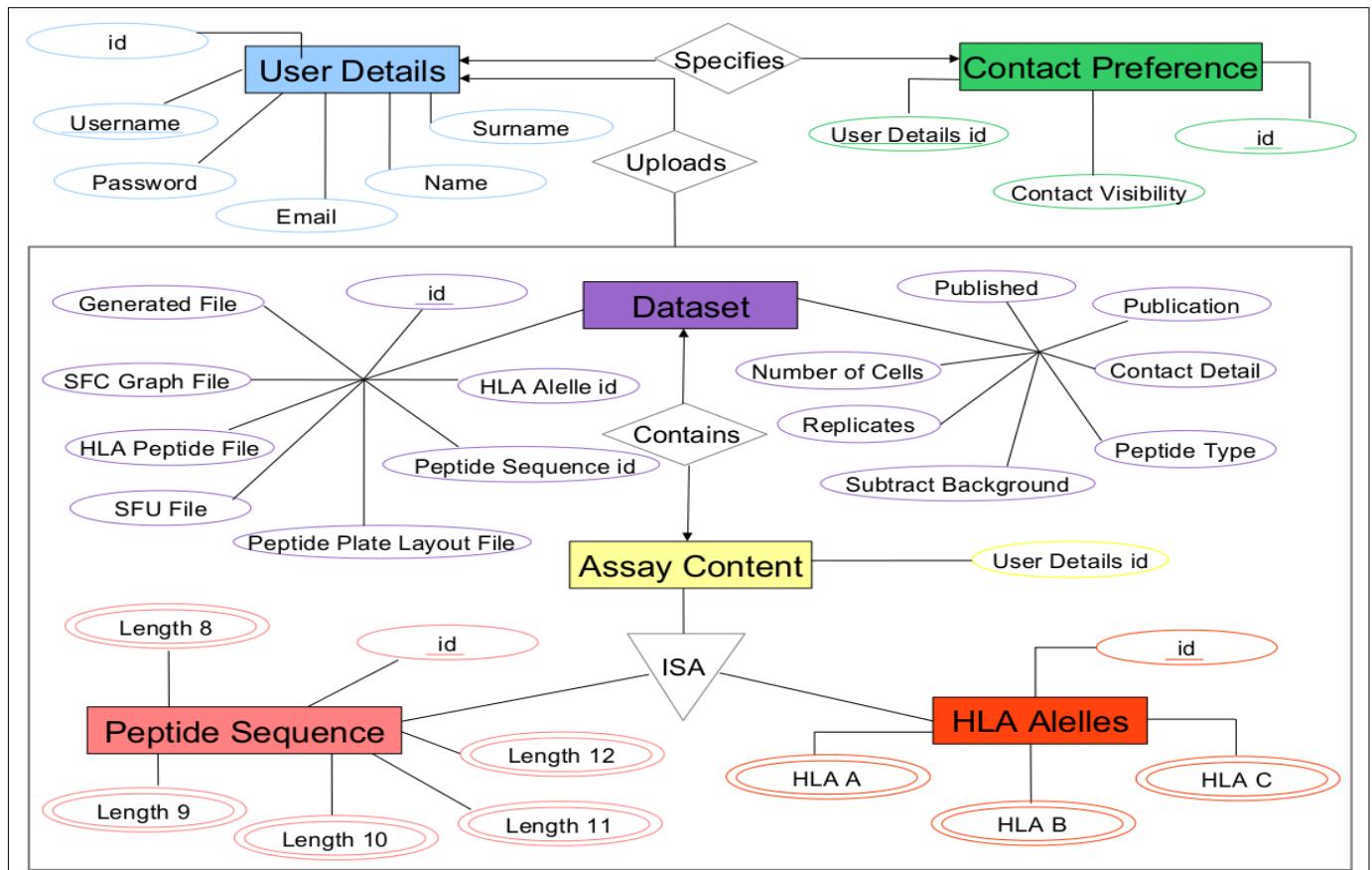


Figure 2.1: An Entity Relationship Model illustrating IMMUNO-SHARE's relational database.

2.2 Evaluating the predictive performance of epitope prediction tools

2.2.1 IFN γ ELISpot assay dataset

Initially, this dataset [80] (obtained from the Los Alamos HIV Immunology database (<http://www.hiv.lanl.gov/content/immunology/hlatem/study1/index.html>)) consisted of 3268 records of reactive overlapping long peptides (OLPs) and corresponding patient identification codes. This dataset was then filtered to ensure that each HLA-allele record consisted of six valid HLA alleles. As a result, 2185 peptide-patient records were used for further analysis. These remaining records were then further sorted according to the HLA alleles they contained into FASTA formatted files contained OLP sequences with each sequence name field containing an associated peptide ID, patient ID, spot forming unit (SFU) value and an automatically calculated SFC value.

Given that the peptides in this dataset were generally 18mers, varying from 15 to 20 amino acids and overlapping by 10 amino acids, a customized version of the Sequence Demarcation Tool - Linux 32 bits (SDT_Linux32; [81]) – as thoroughly explained below in section 2.2.7 - was used to determine which of these overlapping peptides contained overlapping regions.

2.2.2 Selecting four frequently used epitope prediction tools

I selected four frequently used MHC class I epitope prediction tools: netMHC 3.2, IEDEB_ANN, IEDEB_ARB Matrix and IEDEB_SMM (as indicated in Table 1.2). My focus here was to evaluate tools that were open-source, were among the most sensitive and specific, and could be run in a stand-alone mode: a requirement for the comparative tests I wished to perform. All four tools were installed and computations were performed using facilities provided by the University of Cape Town's ICTS High Performance Computing team (<http://hpc.uct.ac.za>).

2.2.2.1 NetMHC 3.2

This tool accepts input epitope sequences in either FASTA or peptide file formats. The latter file format may contain a list of up to 5000 individual peptides, one per line, and all a uniform length (which can be between eight to twelve amino acids long). Prediction of CTL epitopes is restricted to those that bind 78 common HLA molecules where one or more specific HLA molecules may be selected for testing. Binding affinity calculations are based either on weight matrices or ANN prediction values (where values are presented in nanomolar units (nM)) [1].

For ANN predictions, peptides that have associated affinity values less than 50nM are classified as strong binders (SBs), whereas peptides with associated affinity values between 50nM and 500nM are classified as weak binders (WBs). Accordingly, non-binders/non-epitopes (NBs) would be expected to display affinity values greater than 500nM [45].

2.2.2.2 IEDB_ANN

The Immune Epitope Database and Analysis Resource (IEDB-AR) hosts a collection of regularly updated T-cell epitope prediction tools which are specific for MHC class I binding predictions [54]. In addition, IEDB-AR provides access to eight prediction methods (refer to Table 1.2) which have all been re-trained on larger datasets, thus increasing their predictive performance. Amongst these methods, ANN, Average Relative Binding (ARB) and the Stabilized Matrix Method (SMM) are respectively ranked as the top performing methods [1].

I chose to test all three of these methods. The ANN method [45], which is hereafter referred to as IEDB_ANN, accepts input files in FASTA file format, where input sequences may be either a continuous sequence or a list of individual nine amino acid long peptides (each on an individual line). This tool predicts binding between nonameric peptides and a set of 47 different HLA molecules.

Binding affinity calculations of these three methods accessed from IEDB-AR are, as with NetMHC 3.2, based on IC₅₀ values (nM): IC₅₀ values that are less than 50nM indicate SBs and values between 50nM and 500nM indicate intermediate binders (IB). Further, WBs have values between 500nM and 5000nM and NBs values greater than 5000nM [1].

2.2.2.3 IEDB_ARB Matrix

The ARB matrix method is similar to the qualitative binding matrices method, where a matrix of coefficients is constructed from the frequency of each amino acid at any given position in the peptide sequence [50].

This method is only accessible through IEDB-AR and it will hereafter be referred to as IEDB_ARB. It accepts sequences in FASTA format, which may either contain a continuous sequence or a list of individual peptides between eight and eleven amino acids long (each on an individual line and all peptides in one file must be the same length). Predictions are based on a binding affinity between peptides and a set of 56 different HLA molecules with SBs, IBs, WBs and NBs being classified in the same way as IEDB ANN [1].

2.2.2.4 IEDB_SMM

SMM is an improvement on the quantitative binding matrices method where weak binding patterns are detected and data quality is accounted for [51]. Although this method was initially designed to only predict epitopes that bind HLA-A2, an updated version accessible through IEDB-AR can identify epitopes that bind 48 different HLAs and will hereafter be referred to as IEDB_SMM.

All the file input requirements for IEDB_SMM are the same as those for IEDB_ARB and classifications of SBs, IBs, WBs and NBs are made in the same way as for IEDB_ANN and IEDB_ARB [1].

2.2.3 Compiling the testing dataset

To fairly compare the predictive performance of the four tools, the ten HLA alleles which can be analyzed by all the tools were identified – these include A*01:01, A*02:01, A*03:01, A*24:02, A*26:01, B*08:01, B*15:01, B*27:05, B*39:01, B*40:01. As a result, the total number of OLPs within these common HLA allele FASTA format files (provided in Appendix A) are as indicated in Table 2.1.

Table 2.1: List of the ten HLA alleles that are common across the four prediction tools and the total number of OLPs within each of the respective HLA allele files

HLA alleles	No. of OLPs	HLA alleles	No. of OLPs
A*01:01	295	B*08:01	261
A*02:01	940	B*15:01	555
A*03:01	187	B*27:05	64
A*24:02	311	B*39:01	64
A*26:01	228	B*40:01	124

2.2.4. Evaluating the predictive performance of the tools using the testing dataset

The performance of the prediction tools was evaluated for eleven common HLA alleles using only the corresponding HLA allele FASTA format files. In addition, since most HLA molecules have a strong preference for binding nonameric peptides, predictions were made for 9 amino acid long peptides.

Each tool's prediction results were then separated according to the predicted binding type: (i) SB, (ii) IB and WB, and (iii) NB. Within each of these three files, the details of the OLP that contained the specified nonameric peptide were identified. The records in each of these three files were appended to contain the following information: the nonameric peptide, the predicted binding affinity score, the OLP sequence and the OLP ID.

2.2.5. Filtering each tool's prediction results

To begin with, within each of the ten HLA alleles, the nonameric peptide with the strongest binding affinity (i.e. lowest binding affinity) for each OLP was identified. For instance, if at least one of the OLP's nonameric peptides was predicted as a SB then that nonameric peptide was selected. However, if at least one of the OLP's nonameric peptides was predicted as a WB then that nonameric peptide was selected. Furthermore, if all the nonameric peptides of an OLP were predicted as a particular type of binder, then the nonameric peptide with the lowest binding affinity (i.e. strongest binding affinity) was selected.

Thereafter, new datasets were compiled by selecting the nonameric peptide that displayed the strongest binding affinity (i.e. lowest binding affinity) across each patient's corresponding HLA alleles. For example, an OLP was regarded as a SB if at least one of the identified nonameric peptides was predicted as a SB. Similarly, an OLP was regarded as a WB if none of the identified nonameric peptides was predicted as an SB but at least one of the identified nonameric peptides was predicted as a WB.

Moreover, if all the identified nonameric peptides of an OLP were predicted as a particular type of binder, then the lowest binding affinity (i.e. strongest binding affinity) was used in order to categorise the OLP accordingly. As a result, the records in this new dataset displayed the following information: OLP sequence, OLP ID, patient ID, SFU score, automatically calculated SFC score, binding type (SB/WB/NB), nonameric peptide and the corresponding predicted binding affinity score (IC_{50} values (nM)). These pared down datasets consisted of a total number of 1774 records for each of the prediction tools. Most importantly, they are hereafter referred to as my filtered datasets (provided in Appendix B).

2.2.6. Regression analysis of the SFUs and the quantitative predictions of the tools

Firstly, a Spearman's rank-order correlation test (implemented using R - a free software environment for statistical computing and graphing [82]) was used to determine the relative degrees to which SFUs could be predicted based on the inferred binding affinity scores yielded by the four epitope prediction methods.

Lastly, a Spearman's rank-order correlation test was used to determine the relative degrees to which SFUs could be inferred from the strongest predicted binding affinity scores yielded across the four epitope prediction methods.

The null hypothesis, H_0 , for each Spearman rank correlation test is that there is no correlation between the ranks of the predicted binding affinity scores and the ranks of the SFU counts, meaning, as the SFU counts increase, the predicted binding scores should, on average, remain unchanged.

2.2.7 Customized computer scripts

All the computer scripts used in all the analyses conducted are written in python and were run on a Linux 32 bit operating system.

SDT_Linux32 [81] was customized to ensure accurate calculation of the sequence similarity scores for each peptide sequence pair. This modified version firstly calculates the average score of the amino acids present in the respective file where the Block Substitution Matrix 62 (BLOSUM62) score [83] is used to calculate each amino acid's substitution score with itself. Thereafter, all the sequences are aligned, using Muscle [84] with a gap opening penalty of -10 and a gap extension penalty of zero. Lastly, the similarity score for each peptide sequence pair was calculated by using the BLOSUM62 scores and the aforementioned amino acid average scores.

3. RESULTS

3.1 IMMUNO-SHARE, a web-based ELISpot data sharing resource

Despite the potential utility of ELISpot assay data for increasing the power of computational HLA-epitope binding predictions, there is at present no publically available ELISpot data repository. I therefore developed IMMUNO-SHARE, a web-based resource to enable the sharing of ELISpot data by allowing users to easily upload, browse and download ELISpot datasets. (Figure 3.1 displays IMMUNO-SHARE's homepage)



Figure 3.1: IMMUNO-SHARE homepage

3.1.1 System components and features

My main design consideration during the development of IMMUNO-SHARE was the minimisation of demands placed on users. Although it is necessary to create an account in order to upload data to IMMUNO-SHARE, users are able to download data without an account.

3.1.1.1 Uploading procedure

The user-driven addition of data to IMMUNO-SHARE requires two simple steps: (1) the uploading of ELISpot data files and (2) the provision of a small amount of additional information to ensure that the uploaded files are correctly interpreted.

3.1.1.1.1 File uploads

For each uploaded ELISpot experiment, IMMUNO-SHARE expects three different files:

1 A HLA allele, peptide sequence and peptide ID content file. Given that one needs to know the HLA combinations of the patient(s) before conducting an ELISpot assay, each patient's six HLA alleles must be identified and included in this file. In addition, for each peptide that is used in the ELISpot assay, a peptide sequence together with a corresponding peptide identifier must also be included in this file. Consequently, IMMUNO-SHARE will determine the validity of this file by searching through the file's header (the first line of the file) for the following tags: (a) "HLA" (this tag must be detected six times as it will identify the patient's six HLA alleles); (b) "Peptide Sequence", and (c) "Peptide ID". Moreover, as IMMUNO-SHARE is designed to analyze the uploaded files, records within this file that will be selected and used for further analysis should ideally contain the following: (a) six valid HLA alleles identifiers; (b) valid peptide sequences; and (c) a set of suitable peptide identifiers.

2 A peptide ID and plate layout file. On account of 96-well plates being the standard format in which ELISpot assays are conducted, this file should ideally describe the layout of the peptide(s) within the 96-well plate. Firstly, in order to differentiate this file from the uploaded files, IMMUNO-SHARE will search for a hashtag (#). The position of this tag is used to indicate that the adjacent columns and rows are the numerically labelled column and alphabetically labelled row headers of the 96-well plate. Given that there is no standard for the layout of an ELISpot assay, IMMUNO-SHARE seeks to understand each user's specific layout by searching for the positive and negative controls using the tags “PC” and “NC”, respectively. Most importantly, the order of the peptide(s) must be indicated by the peptide identifiers as indicated in file (1). As a result, IMMUNO-SHARE will analyze and validate this file by ensuring that the peptide identifiers in this file correspond to those in file 1. In addition, IMMUNO-SHARE will exclude empty wells by detecting a dash/hyphen (-) tag.

3 A SFU ELISpot reader data file. ELISpot readers are machines used to read and interpret the 96-well plates used during ELISpot assays. These computer-based systems scan ELISpot plates and count the total number of spots in each well - with each spot referred to as a spot forming unit (SFU). IMMUNO-SHARE recognizes this file by detecting the tag “SPOTS NUMBER”. Subsequently, IMMUNO-SHARE will detect the numerically labelled column and alphabetically labelled row headers of the 96-well plate by searching for a #. The unprocessed SFU counts from these files will be used for further analysis. In order to ensure a general file structure of the three requested files, IMMUNO-SHARE accepts excel (.xls) or text (.txt) file formats. Refer to table 3.1 for each file's specific file format(s). Furthermore, examples of the three requested files are attached in Appendix C.

Table 3.1: Upload file specifications: file format restrictions and the required tags in each file

No.	File Name	File Format Restriction ^a	Required Tag ^d
1	HLA alleles, Peptide Sequence and Peptide ID Content	.xls ^b	Peptide ID
			Peptide Sequence
			HLA
2	Peptide ID and Plate Layout	.xls	#
			PC
		.txt ^c	NC
			-
3	SFU ELISpot Reader Data	.xls	Spots Number
		.txt	#

^aThe file format required will be indicated by a file's extension

^b.xls (excel file)

^c.txt (text file)

^dThese tags will assist IMMUNO-SHARE in understanding the file

3.1.1.2 Additional data entry

In addition to the three expected upload files IMMUNO-SHARE will request that a user add a little extra contextual information (Figure 3.2A) so that it can properly interpret the uploaded data. Of primary interest here is the total number of cells in each well. A user can also specify at this point, how many replicates of each sample were examined.

3.1.1.3 Uploading options

In order to make the upload process as easy as possible, IMMUNO-SHARE provides both drag-and-drop and browsing-based upload options:

- 1. Drag and Drop.** As seen in Figure 3.2.B, this uploading option enables the three requested files to be simultaneously uploaded.

2. Browse. For this uploading option, two of the requested files (a) HLA alleles, Peptide Sequence and Peptide ID Content File, and (b) Peptide ID and Plate Layout File) are selected for upload by clicking on the respective browse buttons (as illustrated in Figure 3.3.B). Thereafter, unprocessed SFU data can be copied and pasted directly from the ELISpot reader data file into the space provided (as indicated in Figure 3.3.C).

Upon successful completion of an upload, the submitted dataset is immediately available for download by anybody without any need for an account.

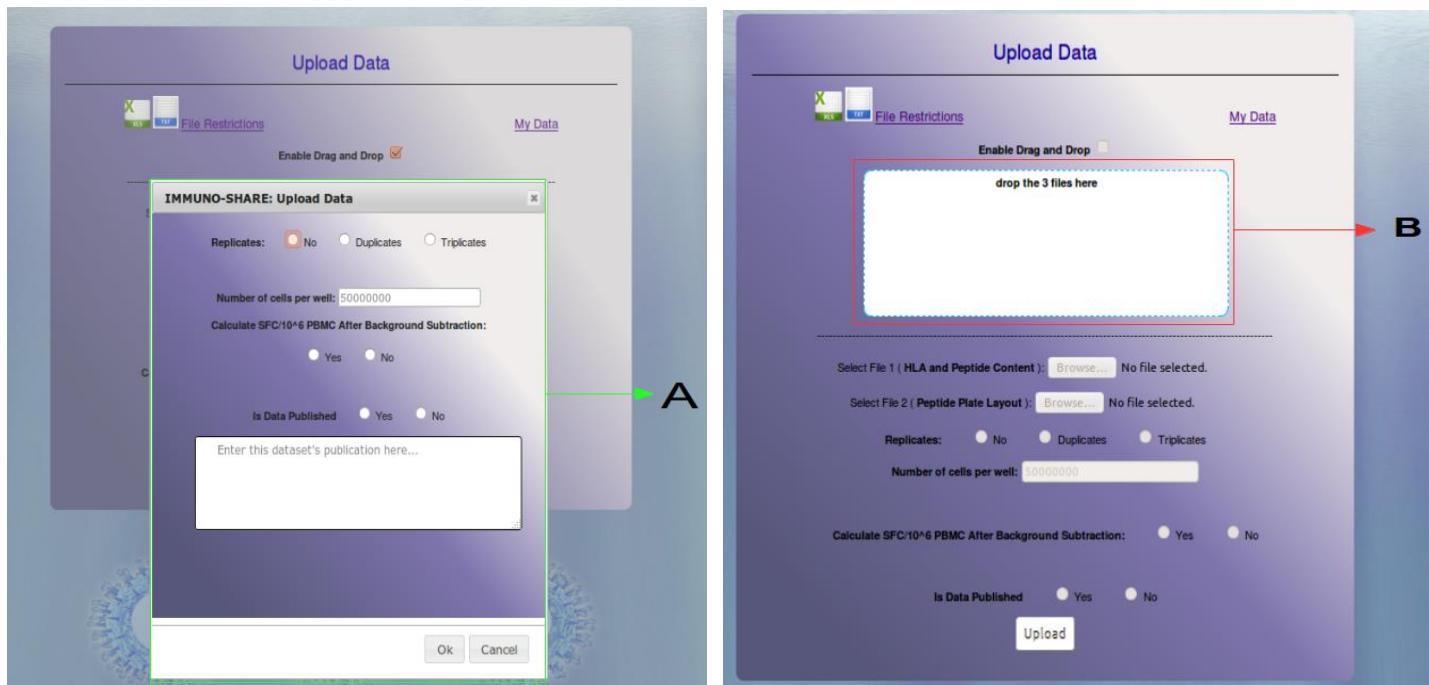


Figure 3.2: These two images show the two sections in the drag and drop uploading option. A. This data entry section provides extra information to help IMMUNO-SHARE understand the implementation of the ELISpot assay. B. This is the drop box for the simultaneous upload of the three requested files.

The screenshot shows the 'Upload Data' page with three main sections highlighted:

- A**: A green box highlights the data entry section containing fields for 'Replicates' (radio buttons for 'No', 'Duplicates', or 'Triplicates'), 'Number of cells per well' (text input set to 5000000), and 'Is Data Published' (radio buttons for 'Yes' or 'No').
- B**: A red box highlights the 'Select File 1 (HLA and Peptide Content)' and 'Select File 2 (Peptide Plate Layout)' browse buttons, both currently showing 'No file selected.'
- C**: A yellow box highlights the large text area for pasting unprocessed SFU data.

Figure 3.3: This image shows the three sections when uploading using the browse option. A. This data entry section provides extra information to help IMMUNO-SHARE interpret how the ELISpot assay was carried out. B. The browse buttons for uploading the two requested files. C. This textbox is reserved for pasting the unprocessed spot forming unit (SFU) data straight from the ELISpot reader data file.

3.1.2 Automatically generated files

With every successful upload the following two files are generated.

(1) **A master file.** First and foremost, as a unique design feature of IMMUNO-SHARE, the requested unprocessed SFUs are automatically adjusted to SFCs per million total cells ($SFC/10^6$). This calculation is based on the number of cells added in a well and the reported unprocessed SFU counts. With regards to any replicates, an average of the unprocessed SFUs is firstly calculated and this average SFU count is adjusted accordingly. Furthermore, IMMUNO-SHARE automatically compiles the relevant information from all three uploaded files into one “master file”. As a result, records within this comma-separated value (.csv) file include the Patient ID, Peptide Sequence, the six HLA alleles, SFU counts

and SFC/ 10^6 counts.

(2) A SFC/ 10^6 bar graph. For a graphical display of the adjusted SFU values from the 96-well plate, IMMUNO-SHARE automatically plots a bar graph image (in portable network graphics (.png) format) using R - a free software environment for statistical computing and graphing [82].

3.1.3 Downloading files

IMMUNO-SHARE users are able to download any of the files within the available datasets. Each dataset consists of the following five files: (1) an HLA allele, peptide sequence and peptide ID content file; (2) a peptide ID and plate layout file; (3) a SFU ELISpot reader data file; (4) a master file; and (5) a SFC/ 10^6 bar graph file.

3.2 Testing the predictive power of epitope detection tools in the context of ELISpot experiments

I attempted to test the hypothesis that four commonly used computational epitope identification tools – netMHC3.2, IEDB_ANN, IEDB_ARB Matrix and IEDB_SMM – were capable of predicting the outcome of ELISpot experiments. If any of these four tools were able to determine the binding affinities of random peptides with even a moderate degree of accuracy, my rationale was that they should be able to predict (to some detectable degree at least) the outcome of ELISpot experiments. Specifically, I used Spearman's rank-order correlation tests to determine whether the estimated epitope binding affinities yielded by each of the four HLA-epitope binding prediction tools had any power to predict observed SFU counts in ELISpot experiments.

As is indicated in Table 3.3, the null hypothesis that the HLA-epitope binding predictions had no power to predict the outcome of ELISpot experiments was only rejected for IEDB_ARB Matrix. However, the positive correlation (as indicated by $rs = 0.052$) is not expected if stronger binding affinities (i.e. lower IC₅₀ values) are associated with higher SFU counts. It is clear therefore that all four of these popular HLA-epitope binding prediction tools performed very poorly in this test with none of the methods yielding predictions of SFU counts that were better than those that could be achieved by random chance.

Table 3.3: Spearman rank test for a correlation between SFU counts and predicted binding affinity scores yielded by the four epitope prediction methods.

Prediction tools	p-value	Rho
netMHC 3.2	0.90	-0.003
IEDB ANN	0.13	0.036
IEDB_ARB Matrix	0.03	0.052
IEDB_SMM	0.21	-0.030

Note: Coefficients printed in bold are significant ($p < .05$).

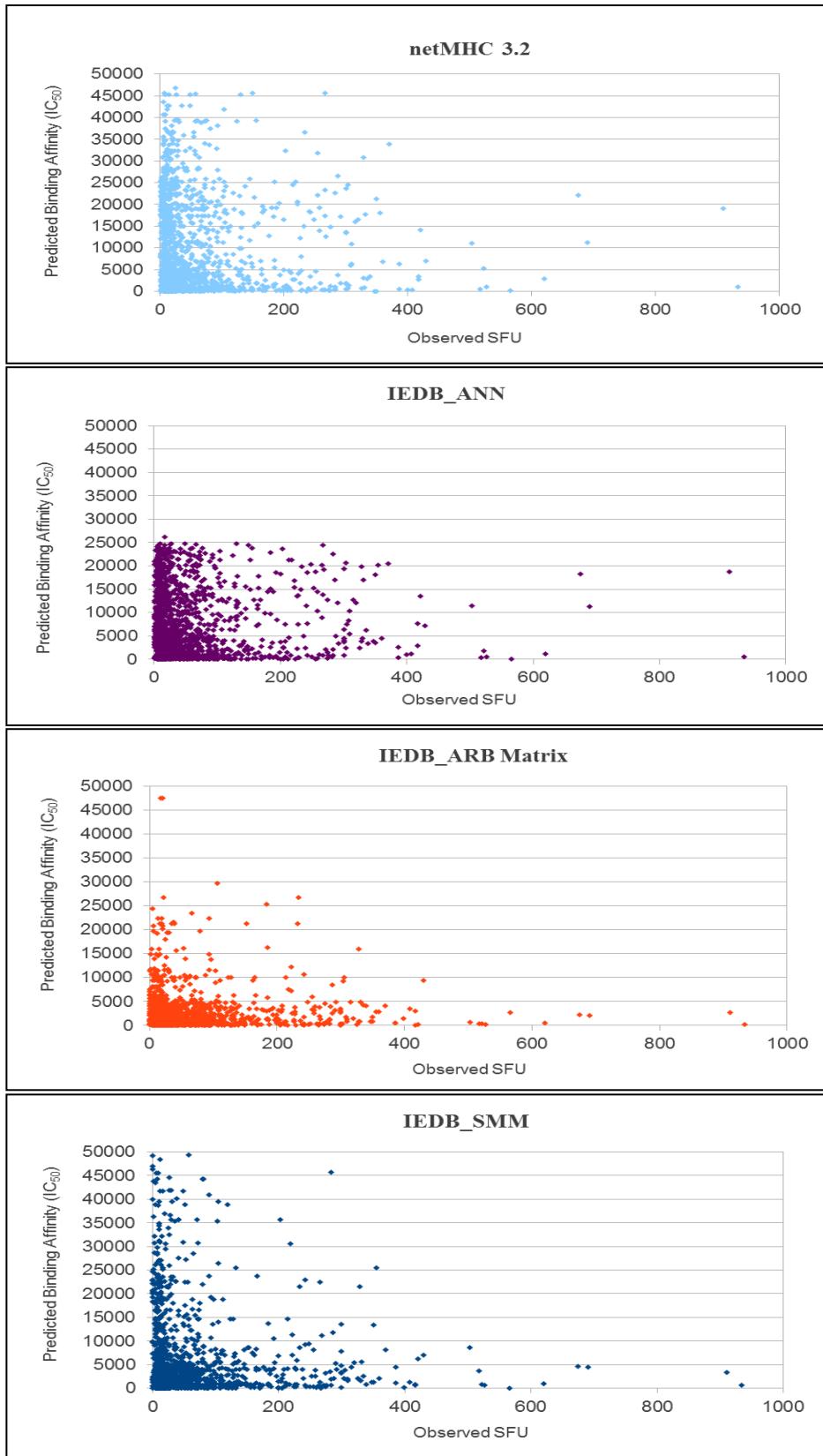


Figure 3.5: Scatter plots graph of observed SFU counts in ELISpot experiments against the predicted binding affinity scores yielded by the four epitope prediction methods, netMHC 3.2, IEDB_ANN, IEDB_ARB Matrix and IEDB_SMM. A spearman rank correlation test failed to reveal significantly negative correlations (i.e. lower binding affinity scores associated with higher SFU counts) for any of these four epitope prediction methods.

3.3 Testing the correlation between the SFUs and the strongest predicted binding affinity scores across the four prediction methods

Given that none of the four tested epitope detection methods was on their own capable of predicting observed SFU counts for common HLA-alleles in ELISpot experiments, I also attempted to determine whether combining the methods would yield an increase in predictive power. This meta-analysis involved testing the correlation between the observed SFUs and the strongest binding affinity (i.e. the lowest binding affinity score) inferred across all four of the tools so as to determine whether these tools might be effectively combined to give more reliable results.

The result of the Spearman's rank-order correlation test carried out on these combined binding affinity estimates yielded a rho of 0.0256 (i.e. a regression line with a positive slope rather than the expected negative slope) and a p-value of 0.281. This test therefore also failed to reject the null hypothesis that there was no association between the binding affinity scores and SFU counts, indicating that even when the results of all four methods are combined they had no power to predict the outcome of ELISpot experiments.

Strongest Predicted Binding Affinity Across The Four Tools

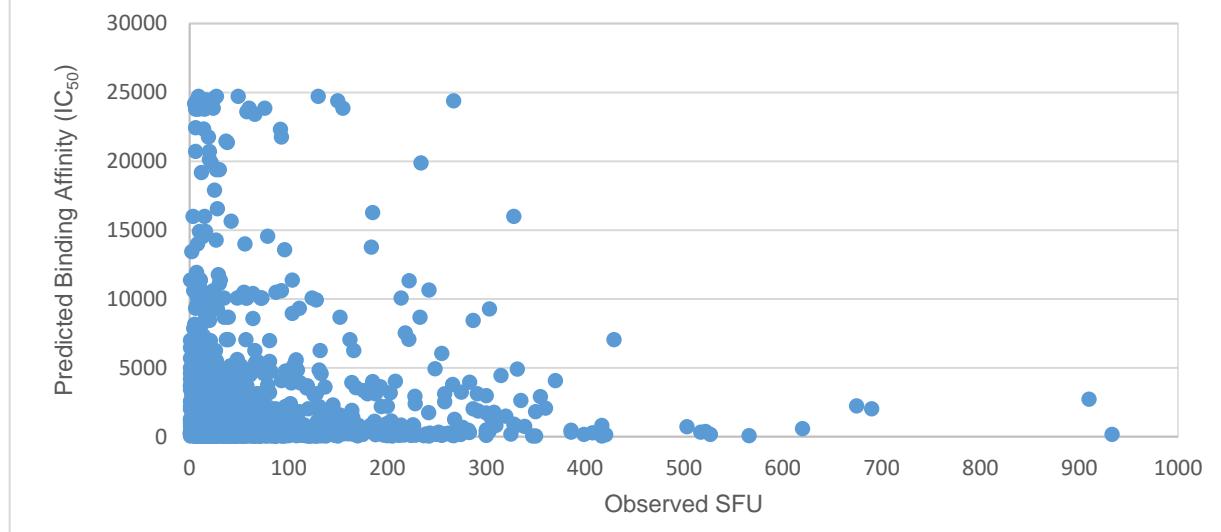


Figure 3.6: Scatter plot of the observed SFU counts in ELISpot experiments against the strongest predicted binding affinity scores (i.e. the lowest score) yielded across the four epitope prediction methods.

4. DISCUSSION AND CONCLUSION

This study explored, (1) and development of IMMUNO-SHARE, a web-driven ELISpot data sharing resource and (2) the use of ELISpot data to evaluate the predictive performance of four frequently used CTL epitope prediction tools (netMHC 3.2, IEDB_ANN, IEDB_ARB Matrix and IEDB_SMM).

Ideally predictors of HLA-epitope binding should be able to accurately infer whether a given peptide sequence is likely to bind to a particular HLA molecule so as to induce an immune response in a patient that would then be detectable in an ELISpot assay. Put another way, in the ELISpot datasets that were analysed here, every overlapping peptide that was associated in these datasets with an immune response in a particular patient should ideally have contained at least one nonameric peptide that was identifiable as a binder to at least one of that patient's HLA alleles. Further, it might be expected that as SFU counts increase (i.e. immune responses become stronger), so too should the predicted strength of binding (i.e. IC₅₀ scores should decrease).

When plotting SFU counts against predicted binding affinity scores, points representing high SFU counts and low predicted binding affinity scores (i.e. true positives) will be situated in quadrant four of the plot (Figure 4.1). Furthermore, points representing low SFU counts and high binding affinity scores (i.e. true negatives) will be situated in quadrant one. Conversely, points in quadrant two will represent SFU counts and predicted binding affinity scores that are both high (i.e. these points indicate false positives) and points in quadrant three will represent SFU counts and predicted binding affinity scores that are both low (i.e. these points indicate false negatives).

Ideally, a good predictor should have a high degree of sensitivity (i.e. yield a high frequency of true positives) and specificity (i.e. yield a high frequency of true negatives) (refer to Table 1.2). Consequently, in Figure 3.5, it is expected that the methods that best predict binding affinities between

particular HLA Alleles and particular peptides should have the highest numbers of plotted points situated in quadrants 1 and 4.

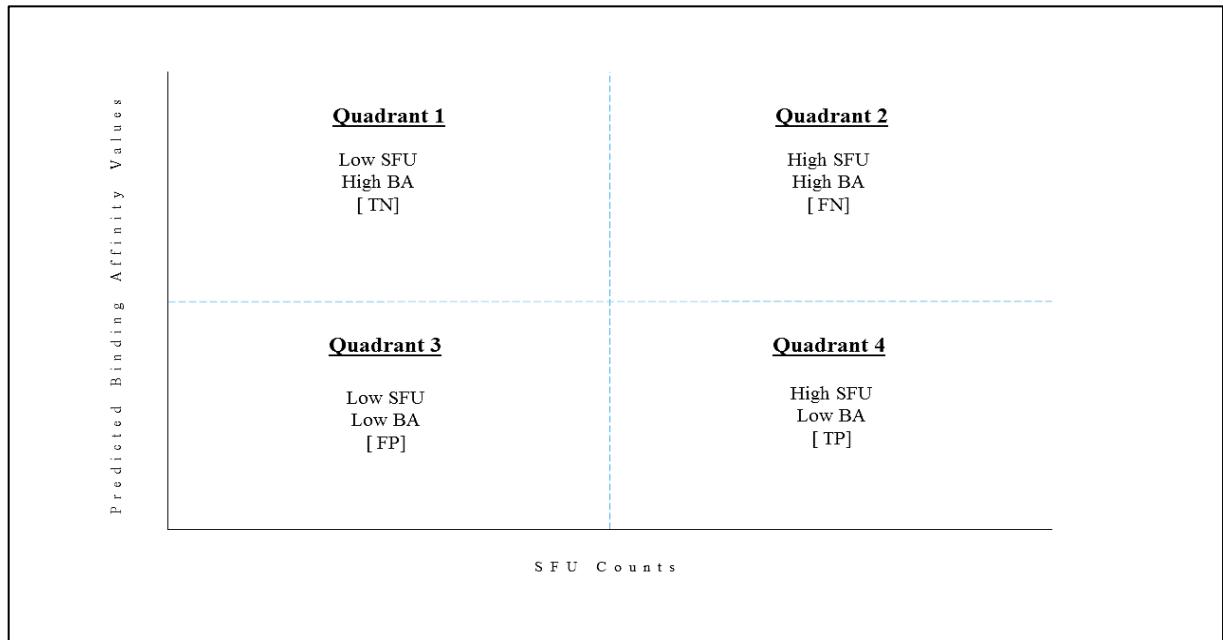


Figure 4.1: An illustration of quadrants within the scatter plots graphs of the observed SFU counts in an ELISpot experiment against the predicted binding affinity scores yielded by the prediction tools represented in Figure 3.5. Points in Figure 3.5 that fall within quadrant 1 are true negatives (TN) in that they represent low SFU counts and low binding affinities (i.e. with high binding affinity scores). Points in Figure 3.5 that fall within quadrant 2 are false negatives (FN), in that they represent both high SFU counts and low predicted binding affinity (i.e. with high binding affinity scores). Points in figure 3.5 that fall in quadrant 3 are false positives (FP) in that they represent both low SFU counts and high predicted binding affinities (i.e. low binding affinity scores). Points in figure 3.5 that fall in quadrant 4 represent true positives (TP) in that they represent high SFU counts and high predicted binding affinities (i.e. low binding affinity scores).

The predictive performances of all four tools was statistically measured using a non-parametric Spearman rank-order correlation test, which indicated that none of the four tested tools yielded binding affinity predictions that were detectably associated with observed ELISpot SFU counts (Table 3.3). As

can be seen in quadrant 3 of the scatter plot graphs of predicted binding affinity scores against observed SFU counts in Figure 3.5, all of the tools yielded high frequencies of false positive binding predictions (i.e. points representing between 0 and 500 SFU counts and below 25 000 predicted binding affinity scores). The low numbers of points in quadrant four of all four scatter plots in Figure 3.5 indicates that all four of the tested methods also yielded only low frequencies of true positive binding predictions (i.e. points representing between 501 and 1000 SFU counts and below 25 000 predicted binding affinity scores). This implies that in order for such tools to have a degree of predictive utility that rivals actual ELISpot-based determination of peptide-induced immune responses, substantial advances need to be made in the quantitative prediction of peptide-HLA interactions that are likely to induce immune responses.

Furthermore, the generally low numbers of true negative binding predictions, represented by low frequencies of points within quadrant one of all four scatter plots in Figure 3.5 (i.e. points representing between 0 and 500 SFU counts and predicted binding affinity scores above 25 000), suggests that, in general, the four tested tools perform very poorly with respect to identifying true negatives – i.e. they have low specificity. This problem is particularly apparent with the netMHC 3.2 and IEDB_SMM methods.

It should be emphasised that the approach that has been used here to test the four epitope prediction tools is especially harsh. The intention of the test that was performed was to determine the divide between what is presently possible with respect to epitope prediction, and what is desirable: a computational approach that will accurately predict the strength of an immune response that is induced by a given peptide within a given individual. In this regard, the four methods that were tested had a significant disadvantage in that peptide binding predictions were made for only a subset of the HLA-alleles that were present within the tested individuals. None of the tested methods are capable of

making peptide binding predictions for the full spectrum of alleles present within these tested individuals. Therefore, the possibility existed that, in many cases, detected SFU counts produced by these individuals in response to particular peptides could have been driven by binding of these peptides to HLA alleles that remained untested. If the tested epitope prediction tools performed perfectly, instances where an untested HLA-allele was responsible for an observed immune response should have yielded low binding affinity predictions for the tested HLA alleles and therefore should have been identifiable as points within the false negative quadrants (quadrant two) of the scatter plots in Figure 3.5. The lack of points within quadrant two of these scatter plots (i.e. points representing between 501 and 1000 SFU counts and above 25 000 predicted binding affinity scores) strongly suggests that the main issue with the four tested analysis tools is not false negative prediction. In the context of the extreme test that these tools have been subjected to, false negative predictions would have, in fact, been a good sign of specificity.

The high false positive rate yielded by the tools analysed in this study – particularly prediction results from IEDB_ARB Matrix and IEDB_SMM – corroborates the findings of another study which evaluated the accuracy of several T-cell epitope prediction tools [76]. However, the predictive performance of all tools in this study - especially the low frequencies of true positive binding predictions by netMHC 3.2 – contradicts the observations of other epitope prediction tool evaluation studies [76, 79]. Although high sensitivity scores were observed in these other studies, it should be pointed out that the tools were tested with datasets containing test peptides that were included within the training datasets of the tools: a factor which means that the apparently high degree of selectivity reported in these studies could have potentially been attributable to over-training and that the results of these studies should therefore be interpreted with caution.

The highly inaccurate predictions made by the four tested tools indicates that users of these tools should not rely on predictions from individual tools. Over the years, users have been advised to combine results from multiple prediction tools in order to obtain greater accuracy, as illustrated by Trost *et al* [79]. However, just as was the case when the results of the four tools were analysed separately, when the results were combined there was still no detectable association between binding affinity predictions and observed ELISpot SFU counts. Given that this meta-analysis approach yielded high frequencies of false positive binding predictions (i.e. points representing between 0 and 500 SFU counts and binding affinity scores below 25 000) and failed to yield high frequencies of false negative binding predictions (i.e. points representing between 501 and 1000 SFU counts and predicted binding affinity scores above 25 000) - as seen in Figure 3.6, it is apparent that there is no guarantee that combining the results from the four analysed tools will substantially increase the reliability of predictions.

As this study indicates, all four of the prediction tools that were analysed would need to improve the accuracy with which they can predict HLA-peptide binding in order for their use to be a reliable alternative to ELISpot experiments. There are a number of ways in which such improvements could be made, including: (1) expanding the repertoire of HLAs that these methods can test (presently only the common HLAs can be tested); (2) the development of new methods within these tools to better identify non-binders (so as to reduce false positives); (3) the development of new meta-tools where multiple different tools are combined to yield higher degrees of specificity and sensitivity than can be achieved with any of the tools individually, and (4) the training of tools using vastly expanded datasets of known HLA-binders and non-binders than have currently been used to train the existing tools. In regard to this latter option, Lin *et al* [76] has suggested that the specificity and sensitivity of these tools could be significantly improved simply by retraining them using larger training datasets. An important potential application of the IMMUNO-SHARE database that has been developed here could be the identification of large numbers of additional HLA-peptide binding interactions. These could then be

used to vastly increase the pool of known HLA-peptide interactions that are presently used to train the binding prediction tools that have been examined here. Besides the quantity of data used to train the tools, another very pertinent factor will be the quality of the binding data. If binding predictions can indeed be made using large volumes of ELISpot data that are placed in the public domain, it will be important to demonstrate that the quality of this data is high enough to enable the accurate definition both of true binders and non-binders and of peptides that will induce immune responses and peptides that will not.

5. FUTURE WORK

Once sufficient IFN γ ELISpot assay datasets are uploaded to IMMUNO-SHARE, it will be possible to directly predict large numbers of peptide-HLA binding pairs that do bind and even larger numbers of pairs that do not. These large datasets of true positive and true negative binders could then be used to massively augment the datasets that have been used to train the present generation of epitope prediction tools.

Further, IMMUNO-SHARE can be modified to incorporate longitudinal IFN γ ELISpot assay data. For instance, IFN γ ELISpot assay datasets containing peptides sampled early in infection and at different points in time (i.e. 12 months, 3 years and etc.) can be linked together. This can provide more information of novel binders and non-binders and, most importantly, will help identify sets of single amino acid substitutions that specifically abolish binding. Such data is especially useful within the context of training and testing datasets for epitope prediction tools since it contains examples of peptides with highly similar sequences but which have completely different binding affinities to particular HLA alleles.

REFERENCES

1. Meraba RL, Ngandu NK, Martin DP. Evaluating the accuracy of Cytotoxic T Lymphocyte epitope prediction tools using known reactive HIV epitopes. Bioinformatics, Honours[thesis], Cape Town: University of Cape Town; 2012.
2. Stevanovic S. Antigen processing is predictable: From genes to T cell epitopes. *Transpl Immunol.* 2005 Aug;14(3-4):171-74.
3. Uebel S, Tampé. Specificity of the proteasome and TAP transport. *Curr Opin Immunol.* 1999 Apr;11(2):203-8.
4. Paulsson KM, Anderson P, Chen S, Sjögren HO, Ljunggren HG, Wang P, Li S. Assembly of tapasin-associated MHC class I in the absence of the transporter associated with antigen processing (TAP). *Int Immunol.* 2001;13(1):23-29.
5. Bacik I, Cox JH, Anderson R, Yewdell JW, Bennink JR. TAP (Transporter Associated with Antigen Processing) - Independent Presentation of Endogenously Synthesized Peptides Is Enhanced by Endoplasmic Reticulum Insertion Sequences Located at the Amino- but not Carboxyl-Terminus of the Peptide. *J. Immunol.* 1994;152:381-7.
6. Sette A, Vitiello A, Reherman B, Fowler P, Nayersina R, Kast WM, et al. The relationship between class I binding affinity and immunogenicity of potential cytotoxic T cell epitopes. *J Immunol.* 1994;153:5586-92.
7. Yewdell JW, Bennink JR. Immunodominance in major histocompatibility complex class I-restricted T lymphocyte responses. *Immuno.* 1999;17:51-88.
8. Pedro AR, Ellis LR. Sequence Variability Analysis of Human Class I and Class II MHC Molecules: Functional and Structural Correlates of Amino Acid Polymorphisms. *J. Mol. Biol.* 2003;331:623-41.
9. Gobin SJP, van Zutphen M, Woltman AM, van den Elsen PJ. Transactivation of classical and nonclassical HLA class 1 genes through the IFN-Stimulated response element. *J Immunol.* 1999; 163:1428-34.
10. Robinson J, Mistry K, McWilliam H, Lopez R, Parham P, Marsh SGE. The IMGT/HLA database. *Nucleic Acids Res.* 2011;D1171-76.
11. Marsh SGE, Albert ED, Boder WF, Bontrop RE, Dupont B, Erlich HA, et al. Nomenclature for factors of the HLA system, *Tissue Antigens.* 2010;25:291-455.
12. Marsh SGE, Albert ED, Boder WF, Bontrop RE, Dupont B, Erlich HA, et al. An update to HLA Nomenclature, *BMT* 2010;45:846-8.
13. Sidney J, Peters B, Frahm N, Brander C, Sette A. HLA class I supertypes: a revised and updated classification. *BMC Immunol.* 2008;9:1.
14. Saito Y, Peterson PA, Matsumura M. Quantitation of peptide anchor residue contributions to class I major histocompatibility complex molecule binding. *J Biol Chem.* 1993 Oct 5; 268(28):21309-17.
15. Naugler C. Origins and relatedness of human leukocyte antigen class I allele supertypes. *Hum Immunol.* 2010;71:837-42.
16. Sidney J, Southwood S, Sette A. Classification of A1- and A24- supertype molecules by analysis of their mhc-peptide binding repertoires. *Immunogenetics.* 2005;57:393-408.
17. Sidney J, Southwood S, Pasquetto V, Sette A. Simultaneous prediction of binding capacity for multiple molecules of the HLA B44 supertype. *J Immunol.* 2003;171:5964-74.
18. Dausset J. The major histocompatibility complex in man-past, present and future concepts. *Science.* 1981;213:1469-74.
19. Altfeld M, Kalife ET, Qi Y, Streeck H, Licherfeld M, Johnston MN. HLA alleles associated with delayed progression to AIDS contribute strongly to the initial CD8+ T cell response against HIV-1. *Plos Medicine.* 2006;3(10):1851-64.
20. McNeil AJ, Yap PL, Gore SM, Brettle RP, McColl M, Wyld R. Association of HLA types A1-B8-

- DR3 and B27 with rapid and slow progression of HIV disease. *Q J Med.* 1996;89:177-85.
- 21. Fellay J, Shianna KV, Ge Dongliang, Colombo Sara, Ledergerber B, Weale M, *et al.* A whole genome association study of major determinants for host control of HIV-1. *Science.* 2007;317:944.
 - 22. Kaslow RA, Carrington M, Apple R, Park L, Muñoz A, Saah AJ. Influence of combinations of human major histocompatibility complex genes on the course of HIV-1 infection. *Nat Med.* 1996;2(4):405-11.
 - 23. Trachtenberg E, Korber B, Sollars C, Kepler TB, Hraber PT, Hayes E. Advantage of rare HLA supertype in HIV disease progression. *2003;9(7):928-35.*
 - 24. Leslie A, Matthews PC, Listgarten J, Carlson JM, Kadie C, Ndung'u T, *et al.* Additive contribution of HLA class I alleles in the immune control of HIV-1 infection. *J. Virol.* 2010 Oct;84(19):9879-88.
 - 25. Kuniholm MH, Kovacs A, Gao X, Xue X, Marti D, Thio CL, *et al.* Specific human leukocyte antigen class I and II alleles associated with hepatitis c virus viremia. *Hepatology.* 2010 May;51(5):1514-22.
 - 26. Borghans JAM, Mølgaard A, de Boer RJ, Keşmir C. HLA alleles associated with slow progression to AIDS truly prefer to present HIV-1 p24. *Plos One.* 2007 Sep;9:1-9.
 - 27. Mendoza D, Royce C, Ruff LE, Ambrozak R, Quigley MF, Dang T. HLA B*5701-positive long-term nonprogressors/elite controllers are not distinguished from progressors by the clonal composition of HIV-specific CD8 + T cells. *J Virol.* 2012;86(7):4014-18.
 - 28. Wilson CC, McKinney D, Anders M, MaWhinney S, Forster J, Crimi C. Development of a DNA vaccine designed to induce cytotoxic t lymphocyte responses to multiple conserved epitopes in HIV-1. *J Immunol* 2003;171:5611-23.
 - 29. Anthony DD, Lehmann PV. T-cell epitope mapping using the ELISPOT approach. *Methods.* 2003;29:260-69.
 - 30. Lehmann PV, Zhang W. Unique strengths of ELISPOT for T cell diagnostics. *Method Mol Cell Biol.* 2012;792(1):3-23.
 - 31. Zhao B, Sakharkar KR, Lim CS, Kanguane P, Sakharkar MK. MHC-Peptide binding prediction for epitope based vaccine design. *J Interg Bio.* 2007;1(2):127-40.
 - 32. Patronov A, Doytchinova I. T-cell epitope vaccine design by immunoinformatics. *Open Biol.* 2013;3:120139.
 - 33. Yang X and Yu X. An introduction to epitope prediction methods and software. *Rev Med Virol.* 2009;19:77-96.
 - 34. Tong JC, Tan TW, Ranganathan S. Methods and protocols for prediction of immunogenic epitopes. *Brief Bioinformatics.* 2006 Oct 13;8(2):96-108.
 - 35. Brusic V, Flower DR. Bioinformatics tools for identifying T-cell epitopes. *Drug Discov Today.* 2004 Jan;2(1):18-23.
 - 36. Martin W, Sbai H, De Groot AS. Bioinformatics tools for identifying class I-restricted epitopes. *Methods.* 2003;29:289-98.
 - 37. Brusic V, Bajic V, Petrovsky N. Computational methods for prediction of T-cell epitopes-a framework for modeling, testing, and applications. *Methods.* 2004;34:436-43.
 - 38. Yu K, Petrovsky N, Schönbach C, Koh JLY, Brusic V. Methods for prediction of peptide binding to MHC molecules: A comparative study. *Mol Med.* 2002; 8(3):137-48.
 - 39. Korber B, LaButte M, Yusim K. Immunoinformatics comes of age. *Plos Comput Biol.* 2006;2(6):0484-92.
 - 40. Peters B, Bui HH, Frankild S, Nielsen M, Lundegaard C, Kostem E, *et al.* A community resource benchmarking predictions of peptide binding to MHC-1 molecules. *Plos Comput Biol.* 2006;2(6):0574-84.
 - 41. Peters B, Tong W, Sidney J, Sette A, Weng Z. Examining the independent binding assumption for binding of peptide epitopes to MHC-I molecules. *Bioinformatics.* 2003;19(14):1765-72.
 - 42. Tsurui H, Takahashi T. Prediction of T-cell epitope. *J Pharmacol Sci.* 2007;105:299-316.

43. Luo H, Ye H, Ng HW, Shi L, Tong W, Mendrick DL, et al. Machine learning methods for predicting HLA peptide binding activity. Bioinformatics and Biology Insights. 2015;9(S3):21-9.
44. Buus S, Lauemøller SL, Worning P, Kesmir C, Frimurer T, Corber S, et al. Sensitive quantitative predictions of peptide-MHC binding by a ‘Query by Committee’ artificial neural network approach. *Tissue Antigens*. 2003;62:378-84.
45. Nielsen M, Lundegaard C, Worning P, Lauemøller SL, Lamberth K, Buus S, et al. Reliable prediction of T-cell epitopes using networks with novel sequence representations. *Protein Sci*. 2003;12:1007-17.
46. Yu K, Petrovsky N, Schönbach C, Koh JLY, Brusic V. Methods for prediction of peptide binding to MHC molecules: a comparative study. *Mol Med* 2002;8(3):137-48.
47. Nielsen M, Lundegaard C, Worning P, Hvid CS, Lamberth K, Buus S, et al. Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach. *Bioinformatics*. 2004; 20(9):1388-97.
48. Buus S, Lauemøller SL, Worning P, Kesmir S, Frimurer T, Corbet S, et al. Sensitive quantitative predictions of peptide-MHC binding by a ‘Query by Committee’ artificial neural network approach. *Tissue Antigens*. 2003;62:378-84.
49. Lundegaard C, Lamberth K, Harndahl M, Buus S, Lund O, Nielsen M. NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11. *Nucleic Acids Res*. 2008;1:36 (Web Server issue):W509-12.
50. Bui HH, Sidney J, Peters B, Sathiamurthy M, Sinichi KA, Purton K, et al. Automated generation and evaluation of specific MHC binding predictive tools: ARB matrix applications. *Immunogenetics*. 2009;57(5):304-14.
51. Peters B, Sette A. Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method. *BMC Bioinformatics* 2005;6:132.
52. Hoof I, Peters B, Sidney J, Pedersen LE, Sette A, Lund O, et al. NetMHCpan, a method for MHC class I binding prediction beyond humans. *Immunogenetics*. 2009;61:1-13.
53. Nielsen1 M, Lundegaard C, Blicher T, Lamberth K, Harndahl M, Justesen S, et al. NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. *Plos One*. 2007;2(8):e796. Available from <http://www.plosone.org> DOI:10.1371/journal.pone.0000796.
54. Vita R, Overton JA, Greenbaum JA, Ponomarenko J, Clark JD, Cantrell JR, et al. The immune epitope database (IEDB) 3.0. *Nucleic Acids Res*. 2015;43:D405-12.
55. Sidney J, Assarsson E, Moore C, Ngo S, Pinilla C, Sette A, et al. Quantitative peptide binding motifs for 19 human and mouse MHC class I molecules derived using positional scanning combinatorial peptide libraries. *Immunome Res*. 2008;4:2 Available from <http://www.immunome-research.com/content/4/1/2> DOI:10.1186/1745-7580-4-2.
56. Kim Y, Sidney J, Pinilla C, Sette A, Peters B. Derivation of an amino acid similarity matrix for peptide:MHC binding and its application as a Bayesian prior. *BMC Bioinformatics*. 2009;10:394. Available from <http://www.biomedcentral.com/1471-2105/10/394>. DOI:10.1186/1471-2105-10-394.
57. Parker KC, Bednarek MA, Coligan JE. Scheme for Ranking Potential HLA-A2 Binding Peptides Based on Independent Binding of Individual Peptide Side-Chains. *J. Immunol*. 1994;152:163.
58. Parker KC, Dibrino M, Hull L, Coligan JE. The μ -microglobulin dissociation rate is an accurate measure of the stability of mhc class I heterotrimers and depends on which peptide is bound. *J. Immunol*. 1992;149:1896-904.
59. Bhasin M, Raghava GPS. Prediction of Promiscuous and High-Affinity Mutated MHC Binders. *Hybridoma and Hybridomics*. 2003;22(4):229-34.
60. Larsen MV, Lundegaard C, Lamberth K, Buus S, Lund O, Nielsen M. Large-scale validation of methods for cytotoxic t-lymphocyte epitope prediction. *BMC Bioinformatics*. 2007;8:424.

61. Larsen MV, Lundsgaard C, Lamberth K, Buus S, Brunak S, Lund O, *et al.* An integrative approach to CTL epitope prediction: A combined algorithm integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions. *Eur. J. Immunol.* 2005;35:2295-303.
62. Nielsen M, Lundsgaard C, Lund O, Keşmir C. The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage. *Immunogenetics*. 2005;57:33-41.
63. Buus S, Lauemøller SL, Worning P, Kesmir C, Frimurer T, Corbet S. Sensitive quantitative predictions of peptide-MHC binding by a ‘Query by Committee’ artificial neural network approach. *Tissue Antigens*. 2003;62:378-84.
64. Peters B, Bulik S, Tampe R, van Endert PM, Holzhütter H-G. Identifying MHC Class I Epitopes by Predicting the TAP Transport Efficiency of Epitope Precursors. *J Immunol.* 2003;171:1741-9.
65. Bhasin M, Raghava GPS. A hybrid approach for predicting promiscuous MHC class I restricted T cell epitopes. *J. Biosci.* 2006;32:31-42.
66. Singh H, Raghava GPS. ProPred1: prediction of promiscuous MHC Class-I binding sites. *Bioinformatics*. 2003;19(8):1009-14.
67. Reche PA, Glutting J-P, Reinherz EL. Prediction of MHC class I binding peptides using profile motifs. *Hum Immunol.* 2002;63:701-9.
68. Dönnes P, Elofsson A. Prediction of MHC class I binding peptides, using SVMHC. *BMC Bioinformatics* 2002;3:25.
69. Dönnes P, Kohlbacher O. SVMHC: a server for prediction of MHC-binding peptides. *Nucleic Acids Res.* 2006;34:W194-7.
70. Rammensee HG, Bachmann J, Emmerich NPN, Bachor OA, Stevanovic S. SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics*. 1999;50:213-9.
71. Dönnes P, Kohlbacher O. Integrated modeling of the major events in the MHC class I antigen processing pathway. *Protein Sci.* 2005;14:2132-40.
72. Adams H-P, Koziol JA. Prediction of binding to MHC class I molecules. *J Immunol Methods*. 1995;185:181-90.
73. Raghava GPS. TMHCpred [Internet]. [place unknown] Available from <http://www.imtech.res.in/raghava/tmhcpred/>.
74. Hakenberg J, Nussbaum A, Schild H, Rammensee H-G, Kuttler C, Holzhutter H-G, *et al.* MAPP – MHC-I antigenic peptide processing prediction. *Appl Bioinformatics*. 2003;2(3):155-8.
75. Trolle T, Nielsen M. NetTepi: an integrated method for the prediction of T cell epitopes. *Immunogenetics*. 2014; Available from DOI:10.1007/s00251-014-0779-0.
76. Lin HH, Ray S, Tongchusak S, Reinherz EL, Brusic V. Evaluation of MHC class I peptide binding prediction servers: Applications for vaccine research. *BMC Immunology* 2008;9:8.
77. Sonego P, Kocsor A, Pongor S. ROC analysis: applications to the classification of biological sequences and 3D structures. *Brief Bioinformatics*. 2008 Jan 11;9(3):198-209.
78. Fawcett T. An introduction to ROC analysis. *Pattern Recogn Lett.* 2006;27:861-74.
79. Trost B, Bickis M, Kusalik A. Strength in numbers: achieving greater accuracy in MHC-I binding prediction by combining the results from multiple prediction tools. *Immunome Res.* 2007;3:5.
80. Frahm N, Korber BT, Adams CM, Szinger JJ, Draenert R, Addo MM, et al. Consistent cytotoxic T lymphocyte targeting of immunodominant regions in human immunodeficiency virus across multiple ethnicities. *J Virol.* 2004 Mar;2187-200.
81. Muhire BM, Varsani A, Martin DP. SDT: A Virus Classification Tool Based on Pairwise Sequence Alignment and Identity Calculation. *PLoS ONE*. 2014;9(9): e108277. Available from doi:10.1371/journal.pone.0108277.
82. R Development Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2011. Available from <http://www.R-project.org/>.
83. Henikoff S, Henikoff JG. Amino Acid Substitution Matrices from Protein Blocks. *PNAS*.

1992;89(22):10915-9.

84. Edgar, Robert C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research. 2004;32(5): 1792-7.