Computational discovery of vaccine targets

Zhang, Guang Lan

2008

Zhang, G. L. (2008). Computational discovery of vaccine targets. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/14139>

<https://doi.org/10.32657/10356/14139>

COMPUTATIONAL DISCOVERY OF VACCINE TARGETS

ZHANG GUANG LAN

School of Computer Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirement for the degree of
Doctor of Philosophy

2008

© Copyright by ZHANG GUANG LAN 2008
All Rights Reserved

III

I certify that I have read this dissertation and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.

KWOH CHEE KEONG
(Principal Advisor)

I certify that I have read this dissertation and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.

VLADIMIR BRUSIC
(Co-Advisor)

I certify that I have read this dissertation and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.

First Reader Name

I certify that I have read this dissertation and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.

Second Reader Name

Approved for the University Committee on Graduate Studies:

IV

Acknowledgements

I would like to express my sincere gratitude and appreciation to my supervisors, Prof. Vladimir Brusic and A/P Kwoh Chee Keong, for their enthusiastic encouragement, excellent guidance and kind support during this research. I would like to thank Prof. Thomas J August from John Hopkins School of Medicine for providing priceless advice. I am also thankful to Dr. K. N. Srinivasan from Health Sciences Authority Singapore, Mr. Asif M. Khan and Ms. Heiny Tan from National University of Singapore, who shared with me their biology knowledge and always gave me their support. It was a great experience working with them. I would like to thank Prof. Wong Lim Soon from National University of Singapore for valuable advices. I appreciate the help of Ms Ivana Bozic from University of Belgrade, in building prediction models based on Support Vector Machines. I am also grateful to Ms. Anitha Veeramani, Mr. Ang Wee Tiong, Dr. Judice Koh, Mr. Justin Choo, Ms. Menaka Rajapakse, Mr. Olivo Miotto, Dr. Paul Tan, Mr. Seng Hong Seah, Mr. Stephanus Daniel Handoko, Mr. Tew Kar Leong, Dr. Victor Tong, Ms. Zhang Zong Hong, and Dr. Zheng Yun for their help and friendship. I am grateful to my parents, my sister and friends for their love and support. Finally, I would like to acknowledge Institute for Infocomm Research and School of Computer Engineering, College of Engineering, Nanyang Technological University for giving me the opportunity to pursue these studies.

Zhang Guang Lan

September, 2007

Singapore

Table of Contents

Acknowledgements	IV
Table of Contents	V
List of Figures.....	VIII
List of Tables.....	XIII
Summary.....	XXI
List of Symbols	XXIII
Chapter 1 Introduction.....	1
1.1. T-cells, MHC molecules, and HLA supertypes	2
1.2. Challenges in Vaccine Targets Identification	5
1.3. Objective of the Thesis.....	7
1.4. Contributions of the Thesis	8
1.5. Organization of the Thesis	10
Chapter 2 Literature Review	12
2.1 Soft Computing.....	12
2.2 Vaccine Discovery	16
2.3 Computational Methods for Identification of T-cell epitopes	18
2.3.1 Structure-based approaches for prediction of MHC binding peptides	19
2.3.2 Sequence-based approaches for prediction of MHC binding peptides.....	21
2.3.3 Amino acid encoding methods	27
2.3.4 Hybrid approaches for prediction of MHC binding peptides	29
2.4 Computational Methods Employed in Our Work	32
2.4.1 Quantitative matrices.....	32
2.4.2 Backpropagation neural networks	33
2.4.3 Hidden Markov models.....	38
2.4.4 Support vector machines	43
2.5 Validation Methods and Performance Measures Used in Our Work	47
2.5.1 Cross validation.....	47
2.5.2 t-test.....	47
2.5.3 Performance measures.....	48
Chapter 3 PRED^{BALB/c}: a system for prediction of peptide binding to the H2^d molecules, a haplotype of the BALB/c mouse	

.....	50
3.1 Introduction	50
3.2 Material and Methods	51
3.2.1 Training datasets and data preprocessing	51
3.2.2 Prediction engine	52
3.2.3 Normalization of prediction scores	52
3.3 Results	53
3.3.1 Ten-fold cross validation	53
3.3.2 Validation of PRED ^{BALB/c} models using independent data	56
3.3.3 Experimental validation of PRED ^{BALB/c} models	58
3.3.4 Performance comparison with other online systems	61
3.4 Discussion	64
Chapter 4 PRED^{TAP}: a system for prediction of peptide binding to the human transporter associated with antigen processing	67
4.1 Introduction	67
4.2 Materials and Methods	69
4.2.1 Training dataset	69
4.2.2 Prediction engines	70
4.3 Results	71
4.3.1 Ten-fold cross validation	72
4.3.2 Model validation using independent datasets	81
4.3.2 Performance comparison to other predictive systems	82
4.4 Discussion	85
Chapter 5 MULTIPRED1: a computational system for prediction of promiscuous HLA binding peptides	88
5.1 Introduction	89
5.2 Materials and Methods	91
5.2.1 Training datasets	91
5.2.2 Data Preprocessing Method	94
5.2.3 Prediction engines	102
5.3 Results	103
5.3.1 Blind testing of HLA-A2 ANN models	104
5.3.2 Blind testing of HLA-A3 ANN models	107
5.3.3 Blind testing of HLA-B7 ANN models	110
5.3.4 Blind testing of HLA-A2 HMM models	112
5.3.5 Blind testing of HLA-A3 HMM models	113
5.3.6 Blind testing of HLA-B7 HMM models	114
5.3.7 Blind testing of HLA-A2 SVM models	115
5.3.8 Blind testing of HLA-A3 SVM models	117

VII

5.3.9	Blind testing of HLA-B7 SVM models	118
5.3.10	Performance comparison of ANN, HMM, and SVM models.....	119
5.3.11	Validation of HLA-A2 prediction models using experimental results.....	123
5.3.12	Validation of HLA-A3 prediction models using experimental results.....	128
5.3.13	Validation of HLA-B7 prediction models using experimental results.....	129
5.3.14	Performance comparison of MULTIPRED1 models with other prediction systems	131
5.4	Discussion.....	138

Chapter 6 Hotspot Hunter: a computational system for large-scale screening and selection of candidate immunological

hotspots in pathogen proteomes.....143

6.1.	Introduction.....	143
6.2.	Material and Methods	145
6.2.1	Training datasets.....	145
6.2.2	Prediction engine	145
6.3.	Model Validation.....	148
6.3.1	Validation of HLA-A2 and A3 prediction models using experimental results	149
6.3.2	Validation of HLA-DR prediction model using experimental results	150
6.3.3	Independent validation of HLA-A2 prediction model.....	154
6.4.	Discussion.....	156

Chapter 7 Contribution and Future Work158

7.1.	Contribution	158
7.2.	Future work.....	162

Author's Publication.....164

Bibliography166

Appendix A User Manual of PRED^{BALB/c}179

Appendix B User Manual of PRED^{TAP}184

Appendix C User Manual of MULTIPRED1.....188

Appendix D User Manual of Hotspot Hunter.....196

VIII

List of Figures

Figure 1.1 MHC class I antigen processing pathway. Figure courtesy of www-bs.informatik.uni-tuebingen.de/Teaching/SS07/CI/Slides/CI_SS07_07.pdf	3
Figure 1.2 Schematic representation of the integrated system.	8
Figure 2.1 A three layer feed-forward network.....	34
Figure 2.2 A Markov chain with five states.	39
Figure 3.1 Sensitivities vs. specificities of the five PRED ^{BALB/c} models obtained from 10-fold cross validation. (A) D ^d , (B) K ^d , (C) L ^d , (D) A ^d , and (E) E ^d	56
Figure 3.2 Sensitivities vs. Specificity for PRED ^{BALB/c} (A) I-A ^d , (B) I-E ^d models for the set of synthetic peptides from eleven proteins with experimentally identified E ^d and A ^d binding.	58
Figure 4.1 Sensitivities vs. specificities of the ANN models in PRED ^{TAP} obtained from 10-fold cross validation. A) H set; B) MH set; C) LMH set.....	75
Figure 4.2 Sensitivities vs. specificities of the HMM models in PRED ^{TAP} obtained from 10-fold cross validation. A) H set; B) MH set; C) LMH set.....	77
Figure 4.3 Sensitivity and specificity of ANN model against thresholds in 10-fold cross validation. The ANN model for prediction of A) H set, B) MH set, and C) LMH set.	78
Figure 4.4 Sensitivity and specificity of HMM model against thresholds in 10-fold cross validation. The HMM model for prediction of A) H set, B) MH set, and C) LMH set.....	80
Figure 5.1 Steps involved in building a prediction system.	89
Figure 5.2 Building the representation of peptide/receptor interaction. The contact residues are those that are involved in interaction. A) Identification of contact residues in the receptor that are involved in peptide/receptor interaction. B) Linear representation of	

IX

peptide and receptor. C) Data presentation used in ANN and SVM model building: Removal of non-contact residues and combining peptide and contact residues in receptor to form a virtual peptide representing peptide/receptor interaction. D) Data presentation used in HMM model building: Removal of non-contact residues and each amino acid in the peptide is followed by its contact residues in receptor to form a virtual peptide representing peptide/receptor interaction.....	95
Figure 5.3 Performance comparison of HLA-A2 ANN models trained using training datasets of various compositions.	106
Figure 5.4 Performance comparison of HLA-A3 ANN models trained using training datasets of various compositions.	109
Figure 5.5 Performance comparison of HLA-B7 ANN models trained using training datasets of various compositions.	111
Figure 5.6 Blind testing performance comparison. A) HLA-A2 models; B) HLA-A3 models; C) HLA-B7 models.	120
Figure 5.7 AROC values of SVM models in blind tests and 10-fold cross-validations for predicting of peptides binding to HLA-A*0201 and *0302.	122
Figure 5.8 Amino acid frequency in each position of 9-mer peptides. A) Peptides binding to A*6801 in Dataset2; B) Peptides non-binding to A*6801 in Dataset2; C) True positive predictions; D) False positive predictions; E) False negative predictions.	138
Figure 6.1 Processing steps in prediction of T-cell epitope hotspots. A) For HLA class I supertypes. B) For HLA class II supertypes.	147
Figure 6.2 A) Predicted HLA-A2 restricted immunological hotspots in HPV type 16 E6 protein. B) Predicted HLA-A3 restricted immunological hotspots in HPV type 16 E6	

X

protein. The red dots represent the start positions of the predicted 30-mer binders and the red bars represent the predicted hotspots. The pink bars represent the positions of experimentally identified hotspots. 152

Figure 6.3 Predicted HLA-A2 restricted immunological hotspots in HPV type 16 E7 protein. The red dots represent the start positions of the predicted 30-mer binders and the red bars represent the predicted hotspots. The pink bars represent the positions of experimentally identified hotspots. 153

Figure 6.4 Predicted HLA-DR restricted immunological hotspots in HCV 1B protein. The red dots represent the start positions of the predicted 15-mer binders and the red bars represent the predicted hotspots. The pink bars represent the positions of experimentally identified hotspots. 153

Figure A.1 An example of the output pages of PRED^{BALB/c} when input is a single protein. The input protein sequence is *Staphylococcal* nuclease (Nase), the sequence type chosen is “protein sequence” and the H2^d alleles of interest are all H2^d alleles. A) The input page. B) The main result page. The input sequence is decomposed into overlapping 9-mers for prediction of binding scores to each allele. C) Graphical view of the predicted binding scores to I-E^d molecule. 182

Figure A.2 An example of the output pages of PRED^{BALB/c} when input is a list of peptides. The input is a list of peptides from chicken ovalbumin and the H2^d alleles of interest are all H2^d alleles. A) The input page. B) The main result page. All 9-mers in each peptide are submitted for prediction. The predicted binding scores are represented by the highest individual 9-mer binding score of each input peptide. C) The alignment view of the predicted binding peptides at threshold 9, which indicates the prediction specificity is 0.9.

XI

.....	183
Figure B.1 The examples of the output pages of PRED ^{TAP} for a single protein. The sequence type chosen is "protein sequence". A) The input page. B) The main result page. The input sequence is decomposed into overlapping 9-mers for prediction of binding scores to TAP. C) Alignment view of the predicted TAP binding regions in the input protein.	186
Figure B.2 An example of the output pages of PRED ^{TAP} for a list of peptides. A) The main result page. All 9-mers in each peptide were submitted for prediction. The predicted binding scores are represented by the highest individual 9-mer binding score of each input peptide. The 9-mer with the highest binding score in each peptide is displayed as "Binding Core" in the result table. B) The sorted result page.....	187
Figure C.1 An example of the output pages of MULTIPRED1 when input is a single protein sequence. The input protein sequence is HPV type 16 E6, the prediction method used is ANN and the HLA supertype of interest is HLA-A2. A) The main result page. The input sequence is truncated into overlapping 9-mers for prediction of binding scores to multiple HLA-A2 alleles, *0201, *0202, *0203, *0204, *0205, *0206, *0207 and *0209. The red ovals are added by authors for clarity of viewing. B) Example graph of Score4. C) Alignment view of the top 4 9-mers in 30-mer windows. D) Prediction hotspot region is 6-47 at threshold 42.	192
Figure C.2 An example of the output pages of MULTIPRED1 when input is a list of peptides. The input protein peptides are from hepatitis C virus, the prediction method used is ANN and the HLA supertype of interest is HLA-A3. A) The main result page. As can be seen here, the input sequence is truncated into overlapping 9-mers for prediction of	

XII

binding scores to multiple HLA-A3 alleles, *0301, *0302, *1101, *1102, *3101, *3301 and *6801. B) Input peptides displayed in the descending order of binding scores. C) Alignment view of the predicted 9-mer binders. 193

Figure C.3 When the model building request is submitted, an intermediate page will be displayed providing the result URL which can be bookmarked for later model retrieval. 194

Figure D.1 An example of the input and output pages of Hotspot Hunter when performing “Single Sequence Query”. The input protein sequence is MOG and the HLA supertype of interest is HLA-DR. A) The input page. B) Prediction result page. C) Plot of predicted scores vs. the amino acid positions in the protein sequence. 200

Figure D.2 An example of the input and output pages of Hotspot Hunter when performing “Multiple Sequences Query”. The input is a set of dengue virus type 1 capsid sequences in ALN (ClustalW) or PHY (PHYLP) format (antigen.i2r.a-star.edu.sg/hh/HTML/alnformat.html) A) The input page. B) Waiting message page. C) Prediction result page. 201

Figure D.3 An example of the output page of Hotspot Hunter when using the “Target Selection” function. The input is a HLA-DR supertype prediction result page for dengue virus type 1 NS2b sequences in HTML format (output of "Multiple Sequences Query" function). 202

List of Tables

Table 2.1 A summary of hybrid methods. (Abbreviations: FL - fuzzy logic, GA - genetic algorithms, ANN - artificial neural networks, HMM - hidden Markov models.).....	15
Table 2.2 A summary of publicly available computational systems related to T-cell epitope prediction.	25
Table 2.3 Amino acid encoding methods.....	28
Table 3.1 Number of 9-mer peptides in the training datasets of PRED ^{BALB/c} class I and class II models.	52
Table 3.2 Sensitivities for PRED ^{BALB/c} (A) A ^d , (B) E ^d models for the set of synthetic peptides from eleven proteins with experimentally identified E ^d and A ^d binding. The values are given for three levels of specificity (0.8, 0.9, and 0.95) for prediction of high (H), moderate (M) and low (L) binders.	57
Table 3.3 The top 5% of PRED ^{BALB/c} predictions vs. T-cell IFN- γ ELISpot peptides in SARS-CoV N. PRED ^{BALB/c} predictions covered all of the nine T-cell IFN- γ immunodominant ELISpot positive sequences from SARS-CoV N. (A) The T-cell IFN- γ ELISpot peptides of SARS-CoV N are indicated by the position number, sequence and PRED ^{BALB/c} alleles predicted. (B) The IFN- γ ELISpot peptides spanning the region N76-114 contained 16 9-mers within the top 5% of predictions of all the five H2 ^d alleles. (C) The IFN- γ ELISpot peptides spanning the region N130-149 contained five 9-mer peptides that were within the top 5% of predictions of four H2 ^d alleles. (D) The IFN- γ ELISpot peptides spanning the region N241-258 contained four 9-mer peptides that were within the top 5% of predictions of A ^d allele. (E) The IFN- γ ELISpot peptides spanning	

XIV

the region N352-370 contained three 9-mer peptides that were within the top 5% of predictions of three H2^d alleles.59

Table 3.4 The top 5% of RANKPEP predictions vs. T-cell IFN- γ ELISpot peptides of SARS-CoV N. RANKPEP predicted seven out of nine T-cell IFN- γ immunodominant ELISpot positive sequences from SARS-CoV N. (A) The T-cell IFN- γ ELISpot peptide sequences of SARS-CoV N are indicated by the position number, sequence and RANKPEP alleles predicted. (B) The IFN- γ ELISpot peptides spanning the region N76-114 contained three 9-mer peptides that were within the top 5% of predictions of K^d and L^d alleles. (C) The IFN- γ ELISpot peptides spanning the region N130-149 contained six 9-mer peptides that were within the top 5% of predictions of K^d, A^d, and E^d alleles. (D) The IFN- γ ELISpot peptides spanning the region N241-258 contained four 9-mer peptides that were within the top 5% of predictions of K^d and A^d alleles. (E) The IFN- γ ELISpot peptides spanning the region N352-370 contained two 9-mer peptides that were within the top 5% of predictions of K^d and E^d alleles.62

Table 3.5 Number of predicted 9-mers which fall in the IFN- γ ELISpot positive regions according to the top 5% of predictions of RANKPEP and PRED^{BALB/c}63

Table 4.1 Number of peptides in the training dataset.70

Table 4.2 Performance assessment of ANN models when using Rep 9 encoding method in 10-fold cross validation.73

Table 4.3 Performance assessment of ANN models when using 5-bit string encoding method in 10-fold cross validation.73

Table 4.4 Performance assessment of ANN models when using conventional sparse encoding method in 10-fold cross validation.....73

XV

Table 4.5 Performance assessment of HMM models in 10-fold cross validation.	74
Table 4.6 Sensitivities and specificities of ANN and HMM models at the selection threshold 6.0.....	74
Table 4.7 Performance assessment of ANN/HMM models when the dataset was partitioned into two parts with the training dataset containing two thirds of the data points randomly selected and the testing set containing the remaining one third of data points. 82	
Table 4.8 The start positions of the top 5% of the predicted TAP binders in HPV type 16 E6 by SVMTAP, TAPPred, and PRED ^{TAP} . The positions highlighted in pink were selected by four prediction models. The positions highlighted in blue were those selected by three prediction models. The experimentally identified HLA-A*0301 binders are ¹ 7–15, ² 33–41, ³ 42–50, ⁴ 59–67, ⁵ 75–83, ⁶ 89–97, ⁷ 93–101, ⁸ 125–133, and ⁹ 143–151. The predictions in the table marked by ^{1–9} are within the 16-mers containing respective HLA-A*0301 binders.83	
Table 4.9 The start positions of the top 2% of the predicted TAP binders in KM-HN-1 by SVMTAP, TAPPred, and PRED ^{TAP} . The positions highlighted in pink were selected by four prediction models. The positions highlighted in blue were those selected by three prediction models. The experimentally identified HLA-A*2402 binders are ¹ 196–204, ² 499–508, and ³ 770–778. The predictions in the table marked by ^{1–3} are within the 16-mers containing respective HLA-A24 binders.	84
Table 5.1 Number of peptides binding or non-binding to 15 HLA-A2 alleles in Dataset1.	92
Table 5.2 Number of peptides binding or non-binding to eight HLA-A3 alleles in Dataset2.	92
Table 5.3 Number of peptides binding or non-binding to 13 HLA-B7 alleles in Dataset3.	

XVI

.....	93
Table 5.4 Peptide positional environment for HLA class I molecules. P denotes peptide residues and R denotes HLA contact residues.	96
Table 5.5 Non-conserved residues of all HLA-A alleles.	99
Table 5.6 Non-conserved residues of all HLA-B alleles.	99
Table 5.7 Non-conserved contact residues across the 15 HLA-A2 alleles appearing in Dataset1. Variations have been highlighted.	100
Table 5.8 Non-conserved contact residues across the eight HLA-A3 alleles appearing in Dataset2. Variations have been highlighted.	100
Table 5.9 Non-conserved contact residues across the 13 HLA-B7 alleles in Dataset3. Variations have been highlighted.	101
Table 5.10 Blind testing on five HLA-A2 alleles: Number of peptides in the training and testing sets for ANN models.	104
Table 5.11 AROC values for blind testing of HLA-A2 ANN models, with four copies of binders in the training datasets.	106
Table 5.12 Blind testing on seven HLA-A3 alleles: Numbers of peptides in training and testing sets for ANN models.	108
Table 5.13 AROC values for blind testing of HLA-A3 ANN models, with four copies of binders in the training datasets.	109
Table 5.14 Blind testing on six HLA-B7 alleles: Number of peptides in the training and testing sets for ANN models.	110
Table 5.15 AROC values for blind testing of HLA-B7 ANN models.	112
Table 5.16 Blind testing of HMM models for five HLA-A2 alleles: Number of peptides in	

XVII

the training and testing sets for HMM models.	113
Table 5.17 AROC values for blind testing of HLA-A2 HMM models.....	113
Table 5.18 Blind testing of HMM models for seven HLA-A3 alleles: Numbers of peptides in training and testing sets for HMM models.	114
Table 5.19 AROC values for blind testing of HLA-A3 HMM models.....	114
Table 5.20 Blind testing of HMM models for six HLA-B7 alleles: Numbers of peptides in training and testing sets for HMM models.	115
Table 5.21 AROC values for blind testing of HLA-B7 HMM models.....	115
Table 5.22 AROC values for blind testing of HLA-A2 SVM models. The three kernel functions are linear, 2-degree polynomial and Gaussian ($g=0.1$ and $C=0.5$).	116
Table 5.23 AROC values and Pearson correlation coefficients for blind testing of HLA-A2 support vector regression models. The kernel functions was Gaussian ($g=0.1$ and $C=0.5$).	116
Table 5.24 AROC values for blind testing of HLA-A3 SVM models. The three kernel functions are linear, 2-degree polynomial and Gaussian ($g=0.1$ and $C=2$).	117
Table 5.25 AROC values for blind testing of HLA-B7 SVM models. Three kernel functions are linear, 2-degree polynomial and Gaussian ($g=0.05$ and $C=0.5$).	118
Table 5.26 Sensitivities in 10-fold cross-validation of SVM, ANN, and HMM models. The values are given for three levels of specificity, 0.8, 0.9, and 0.95. A) HLA-A2 supertype (Dataset1); B) HLA-A3 supertype (Dataset2); C) HLA-B7 supertype (Dataset3).....	121
Table 5.27 HLA-A*0201 binding peptides in HPV type 16 E6 and E7 proteins. “Pos.”, “H”, and “L” stand for the start position of a peptide, high binding affinity, and low	

XVIII

binding affinity.....	124
Table 5.28 HLA-A2 restricted T-cell epitopes from HCMV pp65 and IE-1. The peptides highlighted in yellow are commonly recognized T-cell epitopes. “Pos.” stands for the start position of a T-cell epitope.....	124
Table 5.29 The top 10% of predicted HLA-A*0201 9-mer binders (15 peptides) in HPV E6 in descending order of binding scores. “Pos” stands for the start position of a 9-mer peptide. The bold ones are the true positive predictions. All models picked up the two high binders, but none of them picked up the two weak binders.....	124
Table 5.30 The top 10% of predicted HLA-A*0201 9-mer binders (nine peptides) in HPV E7 in descending order of binding scores. “Pos” stands for the start position of a 9-mer peptide. The highlighted ones are the true positive predictions. ANN models picked up all seven binders.....	125
Table 5.31 The top 5% of predicted HLA-A2 binders in HCMV pp65 (28 entries) in descending order of binding scores. “Pos” stands for the start position of a 9-mer peptide. The peptides highlighted in bold are recognized T-cell epitopes. All three models picked up all T-cell epitopes in pp65.....	126
Table 5.32 The top 5% of predicted HLA-A2 binders in HCMV IE-1 (24 entries) in descending order of binding scores. “Pos” stands for the start position of a 9-mer peptide. The peptides highlighted in bold are recognized T-cell epitopes. ANN and SVM models picked up all T-cell epitopes in IE-1, while HMM model missed out one epitope.	127
Table 5.33 HLA-A*0301 binding peptides in HPV E6 and E7 proteins. “Pos.”, “H”, and “L” stand for the start position of a peptide, high binding affinity, and low binding affinity.	128

XIX

Table 5.34 The top 10% of predicted HLA-A*0301 9-mer binders (15 peptides) in HPV E6 in descending order of binding scores. “Pos” stands for the start position of a 9-mer peptide. The bold ones are the true positive predictions. HMM and SVM models picked up six out of nine binders.....	128
Table 5.35 The top 10% of predicted HLA-A*0301 9-mer binders (nine peptides) in HPV E7 in descending order of binding scores. “Pos” stands for the start position of a 9-mer peptide. The bold ones are the true positive predictions. All models picked up the binder.	129
Table 5.36 The top 2% of predicted HLA-B7 binders (22 peptides) in TRT in descending order of binding scores. “Pos” stands for the start position of a 9-mer peptide. The bold ones are the true positive predictions. ANN models picked up all five binders.	130
Table 5.37 Number of 9-mer peptides in the Peters dataset related to HLA-A2 alleles; number of overlapping peptides between Dataset1 and the Peters dataset; number of non-overlapping peptides (binders/non-binders) used for testing of the HLA-A2 SVM model.	132
Table 5.38 Number of 9-mer peptides in the Peters dataset related to HLA-A3 alleles; number of overlapping peptides between Dataset2 and the Peters dataset; number of non-overlapping peptides (binders/non-binders) used for testing of the HLA-A3 SVM model.	132
Table 5.39 AROC values of HLA-A2 SVM model for predictions of binding peptides for A*0201, A*0202, A*0203, A*0206, A*6802 and A*6901 using the Peters dataset. AROC values of ANN and HMM models are taken from [177]. MULTIPRED ANN model had the highest AROC values in predicting peptides binding to A*0202, *0203, and *0206	

XX

among the external tools evaluated..... 133

Table 5.40 AROC values of HLA-A3 SVM model for prediction of binding peptides for A*0301, A*1101, A*3101, A*3301 and A*6801 using the Peters dataset. AROC values of ANN and HMM models are taken from [177]. MULTIPRED ANN model had the highest AROC values in predicting peptides binding to A*0301 and *1101 among the external tools evaluated. 133

Table 6.1 The six conserved Influenza protein fragments containing multiple T-cell epitopes selected by Stoloff and Caparros-Wanderley..... 155

Table 6.2 HLA-A2 restricted hotspots predicted by Hotspot Hunter in the six Influenza proteins. The six regions shown in Table 6.1 are highlighted in red. A) P35937 – M1A; B) Q71UK7 – M1B; C) P26083 - NPA; D) P04666 - NPB; E) P16509 - PB1; F) P35938 – M2..... 155

Summary

Epitope-based vaccines show great potential in fighting infectious diseases as well as non-communicable diseases. The identification of T-cell epitopes, a crucial step in the design of epitope-based vaccines, is a highly combinatorial problem. Peptide binding to major histocompatibility complex (MHC) molecules is necessary for cellular immune recognition because antigens can only be recognized by T-cells in the form of a peptide complexed by MHC molecules. Experimental approaches for identification of T-cell epitopes are costly, time-consuming, and not applicable to large scale studies. Bioinformatics methods are therefore instrumental for enabling systematic large-scale T-cell epitope mapping. The aim of this work is to aid vaccine targets discovery by combining multiple computational approaches for precise mapping of individual promiscuous T-cell epitopes as well as T-cell epitope hotspots – the regions in protein antigens that have high concentration of these targets.

A suite of computational tools have been developed and validated for the simulation of several important biochemical processes involved in cellular immune responses of human and other organisms. The suite contains a series of related prediction systems: MULTIPRED1 that predicts promiscuous Human Leukocyte Antigen (HLA) binding peptides; PRED^{TAP} that predicts peptide binding to the human transporter associated with antigen processing (TAP); PRED^{BALB/c} that predicts of peptide binding to the H2^d molecules, a haplotype of the BALB/c mouse; and Hotspot Hunter which is for large-scale screening and selection of candidate immunological hotspots in pathogen proteomes. MULTIPRED1 deals with the polymorphism of HLA molecules by focusing on the prediction of promiscuous peptides instead of peptides binding to a single HLA

XXII

molecule. PRED^{TAP} can be integrated with HLA class I binding peptide prediction models for more reliable prediction of T-cell epitopes. PRED^{BALB/c} is the first online server for the prediction of peptides binding to a complete set of MHC molecules in a model organism, providing a more complete view of the immune responses of the organism. Hotspot Hunter is a new generation computational tool aiding in epitope-based vaccine design by accurate identification of immunogenic hotspots in pathogen proteomes through analysis of antigenic diversity.

A variety of classification algorithms including quantitative matrices, artificial neural networks, hidden Markov models, and support vector machines have been assessed and in each case the most suitable solution has been selected and developed. We have also combined the strength of different prediction methods using soft computing approaches to achieve improved performance of prediction systems. The high accuracy and the utility of this suite have been validated theoretically by the author as well as experimentally in independent studies. All the five prediction models in PRED^{BALB/c} and the two models in PRED^{TAP} are both sensitive and specific with AROC (area under the receiver operating characteristic curve) >0.85. The performance of the individual models in MULTIPRED1 are of good accuracy (AROC>0.80). In an independent validation of the models in Hotspot Hunter using six experimentally identified immunogenic peptides, the predicted hotspots fully covered five out of six peptides and partially (80% overlap) covered the last peptide.

List of Symbols

APC	Antigen Presenting Cell
ANN	artificial neural networks
AROC	area under receiver operating characteristic curve
BP	backpropagation
CTL	Cytotoxic T Lymphocyte
ER	endoplasmic reticulum
FL	fuzzy logic
FN	number of false negative predictions
FP	number of false positive predictions.
GA	genetic algorithm
HCMV	human cytomegalovirus
HCV	hepatitis C virus
HLA	Human Leukocyte Antigen
HMM	hidden Markov models
HPV	human papillomavirus
hTAP	human Transporter associated with Antigen Processing
MHC	Major Histocompatibility Complex
MOG	myelin oligodendrocyte glycoprotein
MS	multiple sclerosis
NOD	non-obese diabetic
PR	probabilistic reasoning

XXIV

PBMC Peripheral Blood Mononuclear Cell

SARS Severe Acute Respiratory Syndrome

SE sensitivity

SP specificity

SVM support vector machine

TAP Transporter associated with Antigen Processing

TCR T-cell Receptor

T1DM human type 1 diabetes mellitus

TN number of true negative predictions

TP number of true positive predictions

Chapter 1 Introduction

Amino acids form polymer chains called peptides or proteins. There are twenty naturally-occurring amino acids. Proteins are fundamental components of all living cells that perform specific functions such as enzymes, hormones, and antibodies, among others. Proteins are molecules of life necessary for the proper functioning of an organism. A protein has genetically encoded amino acid sequence which determines its specific shape and function.

Immune system protects an organism from diseases and helps maintain homeostasis – the state of dynamic stability of the organism. There are two types of immune responses, cellular and humoral. Humoral immunity is mediated by B-cells which produce antibody molecules that bind to specific structures on the surface of the pathogen. Pathogens are disease-causing microorganisms, which are broadly recognized in four categories, viruses, bacteria, pathogenic fungi and parasites. Pathogens often are composed of antigenic proteins that stimulate the human immune response. Cellular immunity is mediated by T-cells and is based on the binding of short sequences of amino acids to the binding grooves of specialized cellular proteins known as major histocompatibility complex (MHC) molecules. T-cells are major elements of the immune system of vertebrates and are responsible for generating highly specific and lasting immune responses to pathogenic invaders [1]. The MHC molecules play a vital role in the complex immunological recognition that must occur between T-cells and other cells of the body.

1.1. T-cells, MHC molecules, and HLA supertypes

An antigen is a molecule, generally a protein that is capable of triggering an antibody response in the body. MHC molecules bind short antigenic peptides and transport them to the cell surface where the MHC/peptide complexes are displayed to T-cells of the immune system. T-cells recognize MHC/peptide complexes via T-cell receptors (TCRs) [2-4]. The peptides may be "self", that is originating from the organism's own proteome, or they may be "foreign", originating from bacteria, viruses, environmental antigens, tumors, *etc.* T-cells of a healthy immune system tolerate self peptides while they mount immune responses against foreign peptides. Foreign peptides that provoke immune response are termed immunogenic and they have recognition sites (epitopes) that are recognized by T-cells or antibodies produced by B-cells.

T-cells fall into two major classes, namely CD4⁺ and CD8⁺ T-cells, which differ in the class of MHC molecules they recognize. CD8⁺ T-cells recognize peptides presented by MHC class I molecules that are expressed by all nucleated cells, while CD4⁺ T-cells recognize peptides presented by MHC class II molecules that are presented by the specialized cells of the immune system known as professional antigen presenting cells. The resulting T-cell activation leads to a series of complex events that ultimately lead to T-cell proliferation and the formation of antigen-specific memory T-cells. CD4⁺ helper T-cells release cytokines that regulates the immune response, and CD8⁺ cytotoxic T-cells attack cells containing pathogen proteins.

As shown in Figure 1.1, MHC class I molecules present antigens originating intracellularly. Classical MHC class I antigen processing pathway includes proteasomal cleavage of proteins into shorter peptides, translocation of peptides into the endoplasmic

reticulum (ER) by the transporter associated with antigen processing (TAP), optional ER trimming by amino-endopeptidases, binding of peptides into the binding groove of MHC molecules, and transport of MHC/peptide complexes to the cell surface for presentation to cytotoxic CD8⁺ T-cells (CTLs) [5, 6]. Alternative ways for MHC class I peptide loading exist [7].

MHC class II molecules present extracellular antigens taken up by macrophages or through receptor-mediated endocytosis; e.g. via membrane bound antibodies on B-cells. The MHC class II antigen processing pathway is called endosomal pathway. Efforts have been made for computational modeling of MHC class I antigen processing pathway [8], while MHC class II antigen processing pathway is less studied and to our knowledge, computational systems modeling class II processing pathways are not available.

Figure 1.1 MHC class I antigen processing pathway. Figure courtesy of www-bs.informatik.uni-tuebingen.de/Teaching/SS07/CI/Slides/CI_SS07_07.pdf.

The MHC genes in human, referred to as human leukocyte antigen (HLA), are encoded on human chromosome 6. HLA is organized into three major genetic regions designated class I, II and III. Class I and class II genes encode a number of highly polymorphic cell-surface proteins responsible for antigen presentation. The HLA class I locus has HLA-A, -B, and -C sub-regions. The HLA class II locus, HLA-D, has at least

six sub-regions, namely HLA-DR, -DQ, -DP, -DX, -DO, and -DZ. HLA genes are the most polymorphic human genes known [9].

The HLA alleles that bind overlapping sets of peptides sharing a particular HLA supermotif are collectively referred to as a HLA supertype [10]. Instead of classifying HLA polymorphisms on the basis of serological reactivity, sequence or evolutionary relationships, the classification of HLA molecules into supertypes on the basis of shared structural features and peptide binding motifs could have important practical and theoretical consequences in the development of epitope-based vaccines [11]. Nine major functional supertypes, A1, A2, A3, A24, B7, B27, B44, B58, and B62, covering most of the HLA-A and -B polymorphism, were identified based on largely overlapping peptide binding repertoires and consensus structures in the main peptide binding pockets [12]. Beside experimental approaches for HLA supertype classification, a number of computational methods have been proposed recently. Based on specificity matrices constructed for HLA-A and -B alleles using a nearest neighbor algorithm, Lund and colleagues suggested splitting some of the alleles in the A1 supertype into a new A26 supertype, some of the alleles in the B27 supertype into a new B39 supertype and defining a new B8 supertype [13]. HLA supertypes were also classified using both peptide-derived and protein-derived similarity measures based on learning distance functions [14]. The classification results were in excellent agreement with Sette and Lund's results. It was shown that relatively small collections of supertypes cover a large proportion of the population, for example approximately 88% for A2, A3, and B7 supertypes or 99% for A1, A2, A3, B7, A24, and B44 supertypes in different ethnic groups [12]. Taken individually, each of the HLA class I supertypes A2, A3, B44, and B7

allows coverage of about 35–55% of the general population, irrespective of ethnicity [15].

1.2. Challenges in Vaccine Targets Identification

Peptides that are presented by MHC molecules and recognized by T-cells of the immune system are termed T-cell epitopes. Certain peptides that can be recognized by multiple MHC molecules are called promiscuous T-cell epitopes. They are important targets for vaccine design because they are relevant to a large proportion of human population. Regions of high concentration of promiscuous T-cell epitopes are termed immunological hotspots. They are important for the diagnosis and treatment of infectious, autoimmune, allergic, and neoplastic diseases. The problem of accurate identification of T-cell epitope has been considered a major obstacle to the development of epitope-based vaccines and immunotherapeutics [16].

The binding of the peptides to the MHC molecules is the most selective of the processing steps in the selection of T-cell epitopes. It is estimated that between 1% and 5% of all peptides can bind a particular MHC molecule [17]. Thus, to get a mere 100 independent examples of peptide-MHC binding using a random selection strategy, one would have to test 2000 to 10,000 peptides experimentally. Experimental approaches for identification of T-cell epitopes within these antigens, (i.e. various biochemical and functional assays of overlapping peptides derived from proteins of interest) are costly, time-consuming, and not applicable to studies involving large numbers of protein sequences. Thus, computer models have to be used to help minimize the number of necessary experiments, and enable systematic scanning for candidate T-cell epitopes from larger sets of protein antigens, such as those encoded by complete viral genomes.

The complexity of the vaccine targets identification problem arises from several areas. Firstly, HLA genes show high polymorphism: HLA genes are the most polymorphic human genes known with more than 2700 alleles in the human population identified as of May 2007 [18, 19]. In addition, each human being express three to six HLA class I and at least that number of HLA class II molecules [20]. Because of the high HLA polymorphism, the probability of immune escapes is decreased at the population level as different foreign peptides can be recognized.

Secondly, antigens, especially viral pathogens, show high variation. For example, there are more than 9000 dengue virus (DV) sequences reported in the NCBI Entrez protein database, representing the 10 proteins of the four DV serotypes. In the NCBI Influenza Virus Resource database, there are 10,907 human isolates of influenza A viruses entered from 1997 to 2006.

Finally, for most HLA alleles, there is a relatively small number of experimental data on peptides that bind them. For the vast majority of known HLA alleles, experimental data do not exist at all. Furthermore, imprecision, errors, and biases are a norm in the available experimental data.

Computational vaccinology is in the early stages of development for practical application in vaccine design. The databases of MHC binding peptides [21-25] contain peptide data related to a limited number of HLA variants. The performance of most T-cell epitope or MHC binding peptide prediction systems depends on the amount of available training data and the extent of bias in the dataset [26]. There is a need for the development of accurate and broadly applicable computational prediction systems, which will eventually lead computational vaccinology to become a standard component in

vaccine development for fighting global diseases.

1.3. Objective of the Thesis

This research project focuses on combining multiple software solutions to: a) define new binding rules based on the data available, b) develop methodologies for screening targets of immune responses, c) design improved prediction algorithms by optimization of current algorithms, d) develop more advanced methods to overcome current shortcomings, namely noisy data and lack of experimental data, and e) combine these into a software platform for discovery of vaccine targets. The aim is to help vaccine research by the development of an integrated system for identification of vaccine targets through systematic screening of complete viral genomes. The system needs to contain the following components: a) a viral protein analysis system for selection of representative viral sequences, b) a prediction system that identifies candidate promiscuous MHC binding peptides, and c) a module for interpretation and integration of the prediction results to select the immunogenic regions in the viral proteins.

Prediction of vaccine targets is not a straightforward task due to the complex nature of the biological and medical problem. The function of the systems developed in this project is to speed up design of peptide-based vaccine by providing immunologists with accurate computational predictions of promiscuous MHC binding peptides and immunogenic regions in viral and tumor-associated proteins. The designed system, shown in Figure 1.2, supports vaccine research through identification of the most promising experiments and the reduction of the overall number of required experiments. Specifically, we focus on the selection of key antigenic regions to map promiscuous T-cell epitopes and immunological

hotspots.

Figure 1.2 Schematic representation of the integrated system.

1.4. Contributions of the Thesis

The original contributions of this research work include:

1. Development of PRED^{BALB/c}, the first online server for the prediction of peptide binding to a complete set of MHC molecules in the BALB/c inbred laboratory mouse, a commonly used animal model in immunological studies and vaccine research. The prediction system utilizes quantitative matrices, which were rigorously validated using experimentally determined binders and non-binders and also by *in vivo* studies using viral proteins. PRED^{BALB/c} allows analysis of proteins for the presence of binding motifs to all five MHC molecules in BALB/c mice (H2^d): class I (H2-K^d, H2-L^d, and H2-D^d) and class II (I-A^d and I-E^d) in parallel.

2. Development of PRED^{TAP}, an online system for the prediction of peptide binding to the human TAP. The TAP is a critical component of the MHC class I antigen processing and presentation pathway. The prediction system uses artificial neural networks (ANN) and hidden Markov models (HMM) as predictive engines. PRED^{TAP} can be integrated with prediction systems of MHC class I binding peptides for improved performance of *in silico* prediction of T-cell epitopes.
3. Development of MULTIPRED1, a web-based computational system for the prediction of peptide binding to multiple molecules belonging to HLA class I A2, A3, B7, and class II DR supertypes. The prediction system uses ANN, HMM, and support vector machine (SVM) methods as predictive engines. A data representation method which captures the peptide/HLA interaction information enables MULTIPRED1 to predict peptide binding to multiple HLA alleles with one model per HLA supertype. The main utility of MULTIPRED1 is in the selection of key antigenic peptides to minimize the number of experiments required for mapping of T-cell epitopes.
4. Development of Hotspot Hunter, a web-based computational system for large-scale screening and selection of candidate immunological hotspots in pathogen proteomes through analysis of antigenic diversity. It allows screening and selection of hotspots specific to four common HLA supertypes, namely HLA class I A2, A3, B7, and class II DR. The system uses ANN and SVM methods as predictive engines. Soft computing principles were employed to integrate the prediction results produced by both methods for robust prediction performance. Users can predict hotspots from a single protein sequence, or from a set of aligned

protein sequences representing a pathogen proteome. The latter feature provides a global view of the localizations of the hotspots in the proteome set, enabling analysis of antigenic diversity and shift of hotspots across protein variants. The system also allows the integration of prediction results of the four supertypes for identification of hotspots common across multiple supertypes. The target selection feature of the system shortlists candidate peptide hotspots for the formulation of an epitope-based vaccine that could be effective against multiple variants of the pathogen and applicable to a large proportion of the human population. It is a new generation computational tool aiding in epitope-based vaccine design.

1.5. Organization of the Thesis

This thesis consists of seven chapters. Chapter 1 gives a brief introduction of the biology of the cellular immunity and the bottleneck in vaccine formulation - the identification of T-cell epitopes, and presents the challenges of the research topic as well as objective and contributions of the thesis.

Chapter 2 provides a literature survey covering soft computing methods and their applications in bioinformatics, vaccine discovery, computational methods and prediction systems aiding T-cell epitope discovery. The four computing methods employed in this work, quantitative matrices, ANN, HMM, and SVM, are introduced in detail. The performance measures used in this work are introduced too.

Chapter 3 describes $\text{PRED}^{\text{BALB/c}}$, an online system for the prediction of peptide binding to the H2^{d} molecules, a haplotype of the BALB/c mouse. The training datasets, the testing methods, and validation results are reported. The performance comparison

between PRED^{BALB/c} and RANKPEP, a publicly available prediction server, is reported too. The user manual of PRED^{BALB/c} is presented in Appendix A.

PRED^{TAP}, a web-based system for the prediction of peptide binding to the human TAP, is reported in Chapter 4. The prediction methods applied, ANN and HMM, are introduced. The training dataset, the model testing methods, and validation results are reported. The performance comparison of PRED^{TAP} and several publicly available prediction servers, SVM-TAP and TAPPred, is reported. The user manual of PRED^{TAP} is presented in Appendix B.

Chapter 5 introduces MULTIPRED1, a web-based computational system for prediction of promiscuous HLA binding peptides. A novel data preprocessing method which captures the peptide/MHC interaction information, methods for model building and testing, and model optimization techniques are described. The prediction validation results and performance comparison with some publicly available MHC binding peptide prediction servers are reported. The user manual of MULTIPRED1 is presented in Appendix C.

Chapter 6 reports Hotspot Hunter, a computational system for large-scale screening and selection of candidate immunological hotspots in pathogen proteomes. The soft computing method used in combining the ANN and SVM models is described. The prediction models were validated using experimental data and the validation results are reported. The user manual of Hotspot Hunter is presented in Appendix D.

Chapter 7 summaries the contribution of the thesis and discusses future research directions.

Chapter 2 Literature Review

Immunoinformatics is the application of computational technology to the study of immunologically important processes [20]. It facilitates the understanding of immune function by modeling the interactions among immunological components. In this chapter, the two major aspects of immunoinformatics, the computational technology and the immunological processes, are reviewed.

2.1 Soft Computing

The concept of soft computing was introduced by Zadeh [27]. It refers to a collection of computational techniques in computer science, artificial intelligence, machine learning and some engineering disciplines, which attempt to study, model, and analyze very complex phenomena. In contrast to traditional hard computing, soft computing tolerates imprecision, uncertainty and partial truth [28]. The constituents of soft computing offer their own advantages and techniques in partnerships to allow solutions to otherwise unsolvable problems. They are complementary, not competitive to one another. Soft computing is causing a breakthrough in the engineering and scientific fields since it helps solve problems that could not be solved by traditional analytical methods [29]. Key areas of soft computing include fuzzy logic, evolutionary algorithms, neuro-computing, and ideas about probability.

Fuzzy logic (FL) [30] is a problem-solving control system methodology that provides a simple way to arrive at a definite conclusion based upon vague, ambiguous, imprecise, noisy, or missing input information. FL mimics human decision making; translates

qualitative knowledge into a mathematical language and provides formalism for problem solving. It provides a technique to deal with imprecision and poorly defined information granularity. Its main characteristic is the robustness of its interpolative reasoning mechanism.

Genetic algorithms (GA) are the most popular type of evolutionary algorithms. They are inspired by the biological evolution process [31]. The aim is to produce a new generation which is a better solution to the problem than the current generation, and three basic operations are used in pursuit of this goal. Firstly, the members of the population are ranked according to their 'fitness' (which is determined for a member based on its suitability as a solution), and members chosen for mating are selected with a probability proportional to their fitness. This is called selection. Secondly, a technique called *crossover* is applied to selected members of the population to maximize retention of the good points of the previous generation. This is analogous to biological mating, in which a child may be superior to both parents if it inherits 'good genes' from both of them. Thirdly, random mutations are made on some individuals of the new generation. The new population generated by these operations replaces the old population. GA gives us a way to perform randomized global search in a solution space.

Artificial Neural networks (ANNs) [32] are non-linear statistical data modeling tools often used to model complex relationships between inputs and outputs. ANNs can mimic the human's ability to adapt to changing circumstances; this depends heavily on learning from events that have happened in the past and applying this knowledge to future situations. ANNs consist of many nodes which are processing units analogous to neurons in the brain. Given an input, the output is determined by the node activation functions

coupled with a set of local parameters. Information is “learned” and distributed into an ANN structure. A basic architecture of an ANN has an input layer, one or more hidden layers and an output layer, all connected through weighted links. The procedure used to perform the learning process is called a learning algorithm. A learning algorithm determines the procedure for modification of the weights so that an ANN can learn generalized patterns from the training data.

Probability theory is the branch of mathematics concerned with the analysis of random phenomena. Hidden Markov Models (HMM) are a concept in probability theory widely known for their application in the speech recognition. They have gained increasing acceptance in bioinformatics field as means of sequence modeling, multiple alignment and profiling[33]. In this work, an HMM has been employed in prediction of TAP binding peptides and MHC binding peptides. Another method belonging to probability ideas used in this work is position specific probabilistic matrices, which has been used for prediction of MHC binding peptides.

In addition to these techniques, there are a large number of hybrid algorithms, in which two or more techniques were integrated to complement each other to achieve overall better performance. A standard ANN learning algorithm uses the backpropagation (BP) algorithm which is based on the gradient descent principle. It usually converges faster than alternative methods, such as GA. However, the performance of BP is dependent on the correct choice of the parameters – learning rate, momentum and the initial values of the net’s weights and thresholds. It often suffers from getting trapped in a local minimum which may result in non-convergence during training of the network, or may perform poorly. GAs have the advantage of being robust – they usually do not get stuck in local

optima, however they may take a long time to converge because the search space is much larger [34, 35]. A hybrid GA-ANN algorithm was proposed by [36], in which the GA finds a location in the parameter space at which the value of the error function is below a certain threshold, and then uses the values at this location to initialize the ANN instead of random values. McInerney and Dhawan used a different way to combine GA and BP, in which the GA is used to escape from any local minima that BP may fall into as opposed to using GA to find a good launching location for BP [34, 35]. They explored two forms of hybrid GA-BP algorithm: GA-BPbasic, which requires the momentum and learning rate parameters for BP, and GA-BPadap, which adapts the learning rate as training continues and does not require the momentum or learning rate be specified. Both hybrid forms demonstrated superior performance to standard BP neural networks. Table 2.1 lists a summary of hybrid methods.

Table 2.1 A summary of hybrid methods. (Abbreviations: FL - fuzzy logic, GA - genetic algorithms, ANN - artificial neural networks, HMM - hidden Markov models.)

Technologies	Characteristics
ANN controlled by FL	FL controls the learning rate of ANNs. It provides automatic tuning of the learning rate parameter depending on the shape of the error surface. It dramatically improves training time for some problems [37].
GA controlled by FL	FL was applied to manage the resource of GAs during their global search in the solution space to localize the search in the discovered promising regions of that space. This improves training time of GA [38].
FL tuned by GA	GA was used to modify the membership functions in the term sets of the variables used by the FLs. It provides the system with adaptive capabilities [39].
ANN tuned by GA	There are many ways for GAs to tune ANNs, such as finding the optimal set of weights for a given topology. Combines advantage of ANN and GA. It is more robust than ANN and converges faster than GA [36].
FL tuned by ANN	Many combinations of FL and ANN exist. ANN and FL are not only similar, but also complementary to each other. FL deals with explicit

	logic by describing it in rules, while ANNs have the learning capability that FLs do not have. ANN+FL fused models provide both advantages [40, 41].
Hybrid GA	GA and Hill Climbing (HC) techniques were integrated. There are two ways of integration: one is to embed the HC in the crossover operator of GA, the other one is to embed the HC in the solution selection step [42].
HMM controlled by GA	GA was used to optimize the structure of HMM. Automatic optimization of the structure of HMMs was achieved [43].
Probabilistic matrix derived using GA	GA was used to generate probabilistic matrices. The matrix generated achieved higher generalization ability [44].

Recently, soft computing approaches have been effectively applied in some bioinformatics problems in which the data is of low quality. For example, neuro-fuzzy methods have been applied in knowledge discovery in databases where algorithms able to deal with uncertainty and imprecision are needed [45]. A neuro-fuzzy based approach has also been used for virtual screening of active ligands that bind target molecules [46]. A fast leave-one-out evaluation-guided gene selection algorithm was also applied for microarray data analysis with least-squares SVMs [47].

2.2 Vaccine Discovery

Vaccine is a suspension of attenuated or killed bacteria or viruses, administered for the prevention, amelioration or treatment of infectious diseases. Vaccines help prevent morbidity and mortality due to infectious diseases and they have been among the greatest public health successes. The global eradication of smallpox declared in 1980 is considered as one of the most significant public health achievements of all time [48]. Recently, vaccine technology has also found increasing applications in control of

non-communicable diseases, such as cancers, allergy, autoimmunity and even dementia [49-52].

The science of vaccination began just two centuries ago with scientific studies of prevention of smallpox through inoculation with cowpox by Jenner and the modern era vaccinology began in mid-twentieth century [53]. Currently, there are more than 30 approved vaccines for human, around half of which are commonly used, mainly to prevent childhood infections [48]. Vaccine research has grown to be an important research field with global annual R&D funding of around US\$5 billion [54].

However, even with the past successes and all the research effort, infectious diseases, such as tuberculosis, malaria, and HIV/AIDS, continue to cause around 25% of global deaths [55]. Every year, some three million people die of tuberculosis, and one million of malaria. More than 20 million people have died from HIV/AIDS, which is the leading cause of death for young adults worldwide, and an estimated 34 to 46 million others are now infected with the virus [56]. Moreover, the levels of resistance to antibiotics by many viruses and bacteria are increasing, for example, the constant possibility of a dengue fever outbreak and a major influenza pandemic. Thus, there is still urgent need to design and develop vaccines to fight against these diseases.

Most traditional vaccines uses whole pathogens, but the use of macromolecules is becoming more and more important as using only a defined part of the pathogen (epitope) is safer and usually has fewer side effects. Studies showed that epitope-based vaccines are promising measures in the fight against HIV1, malaria, and cancers [57-59]. A first step of the epitope-base vaccine strategy is the identification and selection of T-cell or B-cell epitopes that act as vaccine targets.

2.3 Computational Methods for Identification of T-cell epitopes

The development of computer aided vaccine design, or computational vaccinology, is one of the major goals of immunoinformatics [60]. The key to solving this challenge is the prediction of immunogenicity; in other words, it is to predict which snippets of a protein, out of hundreds or sometimes even thousands of possible candidates, are most likely to elicit a strong immune response.

The identification of T-cell epitopes is a crucial step in the design of epitope-based vaccines. The binding of the peptides to the MHC complex molecules is the most selective of the processing steps in the selection of T-cell epitopes. Thus, the *in silico* prediction of immunogenicity with focus on the prediction of peptide binding to MHC, is a necessary step in systematic identification of T-cell epitopes [60]. The interaction between MHC molecules and peptides usually involves a sequence of nine amino acids spanning the MHC groove. For example, HLA class I molecules interact with the peptide N- and C-termini, imposing a typical bulge on the peptide chain and allowing only peptides of a defined length, typically 8-10 amino acids, to be accepted for binding [61]. The binding groove in HLA class II molecules has open ends and peptides are bound in an extended conformation, allowing the peptides to be more variable in length (typically, 12-20 amino acids) [62]. HLA class II binding peptides have a core binding region, nine amino acid long, that fit into key pockets in the HLA groove, accommodating peptide side chains in a similar fashion to the class I alleles [63]. However, it is important to note that efficient peptide/MHC binding does not guarantee immunogenicity [64] and binding to MHC alleles is only a prerequisite for peptides being T-cell epitopes.

The problem of predicting peptides that bind to MHC molecules has been extensively

investigated since the late 1980s. The computational methods fall in two major categories, structure-based and sequence-based approaches. Structure-based methods are especially useful when no experimental peptide binding data are available. They are computationally intensive, and have mainly been applied to MHC molecules with known crystal structures. Structure-based approaches include protein threading, homology modeling, molecular dynamics and docking *etc.* Sequence-based approaches rely on the primary sequence of peptides known to experimentally bind to a target MHC molecule. Because they are computationally efficient, most of the publicly available computational tools for prediction of MHC binding peptides employ sequence-based methods. Sequence-based approaches include binding motifs, quantitative matrices, decision trees, ANN, HMM, SVM, and ensemble classifiers. The publicly available prediction tools for MHC binding peptides are summarized in Table 2.2.

2.3.1 Structure-based approaches for prediction of MHC binding peptides

Structure-based approaches rely on the structural conservation observed in crystallographically solved peptide-MHC complexes. Molecular dynamics simulation, one of the first structure-based computational approaches, models peptide binding by calculating the forces exerted on every atom of MHC-peptide complexes [65-69].

Protein threading methods evaluate the compatibility of a particular protein sequence with a protein structure [70]. In the MHC binding problem, peptide candidates are threaded using the peptide structure in the MHC groove as a template, and their binding compatibility is evaluated by statistical pairwise potentials [71, 72]. These methods are independent of MHC binding data, but the structure of the MHC-peptides complex needs

to be available. PREDEP, an online prediction server, employed protein threading for prediction of MHC class I binding peptides [72-74].

Homology modeling, also called comparative protein modeling, derives the most probable structure for a sequence given its alignment with related structures [75, 76]. In the case of peptide/MHC prediction, the bound conformation of a peptide sequence with an unknown structure is modeled based on the structure of other bound peptides to homologous MHC molecules. EpiDock integrates homology modeling with a free energy scoring function FRESNO for prediction of MHC class I binding peptides [77]. First, the 3-D structures of all possible MHC-peptide complexes are obtained by homology modeling. Then a free energy scoring function is used to predict the absolute binding free energy of all possible peptides to the MHC class I molecule. The method only requires a minimum of five experimentally-determined binding free energies to derive a predictive empirical scoring function.

Docking is a useful technique for investigating intermolecular interactions. The docking simulation finds the most probable translational, rotational, and conformational juxtaposition of a given ligand-receptor pair, and calculates how well a ligand can bind to the receptor. Several docking techniques have been used to study the binding of peptides to MHC molecules [78-82].

A computational combinatorial design approach has been used for predicting sequences and structures of the peptides that bind an MHC molecule of known X-ray crystallographic or NMR-determined structure [83]. Three-dimensional quantitative structure-affinity relationship (3D QSAR) studies using comparative molecular field analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSIA) have

been employed too [84]. PredMHC, an online server, has been developed based on 2D QSAR approach to predict peptides binding to HLA class I molecules [85, 86]. The updated version of the server, PredMHC 2.0, includes prediction models for peptides binding to 11 HLA class I, three HLA class II, three mouse class I molecules and TAP [87]. Recently, an algorithm, PePSSI (peptide-MHC prediction of structure through solvated interfaces), was proposed for the prediction of peptides binding to HLA-A2 molecules [88]. It integrates sampling of peptide backbone conformations and flexible movement of MHC side chains and is unique in its incorporation of water molecules at the peptide-MHC interface.

2.3.2 Sequence-based approaches for prediction of MHC binding peptides

A complementary sequence-based approach is based on information provided in experimental peptide binding data. Binding motifs, the earliest methods used for identification of MHC binding peptides, describe sequence patterns usually extracted from large numbers of existing known MHC binding peptides, or from pool sequencing experiments [89]. They indicate which positions and which amino acids are important for binding. SYFPEITHI, one of the earliest publicly available online MHC databanks, has data on MHC binding peptides and peptide motifs [22]. It also serves as a prediction server for MHC binding peptides using the binding motif method. The binding motif method has been widely applied for MHC binding peptide prediction [90-92]. It was found that besides anchor positions, other positions in a peptide and between-position interactions also play a role in binding to MHC molecules. However, peptides that do not conform to binding motifs were identified to be binding peptides [93, 94]. Motif-based

methods are of relatively low accuracy – it has been estimated that approximately a third of the peptides predicted as binders to a specific MHC molecule actually bind [95].

Quantitative matrices provide more detailed models in which the contribution to binding of each amino acid at each position within a binding core of a peptide is quantified [96, 97]. The position-specific amino acid values reflect the structural properties of HLA alleles, therefore constituting a sort of “fingerprint” of HLA binding domains. Quantitative matrix-based prediction systems are linear models that are easy to implement, and that result in a binding score for each query peptide. Their predictive accuracy is considerable, but they have less capacity to encode non-linear dependencies. Several online prediction systems, such as BIMAS [97], EPIPREDICT [98], and PRED^{NOD} [99] are based on quantitative matrices. PRED^{NOD} is an online prediction system for identification of peptides that bind the MHC molecules constituting the H-2^{g7} haplotype. The H-2^{g7} haplotype comprises the MHC class-I molecules K^d and D^b and a single class-II molecule I-A^{g7}. The derivation of a scoring matrix for I-A^{g7} was difficult due to the small number of experimentally validated non-binders and inconsistencies between experimentally determined peptide datasets determined in various studies. This difficulty was resolved by using an evolutionary algorithm and producing a binding matrix that predicts well across disparate datasets [44]. RANKPEP [100, 101] uses position specific scoring matrices (PSSM) for MHC binding prediction and it predicts proteasomal cleavage sites too. Stabilized matrix method (SMM) is an improved matrix-based algorithm combining a matrix with pair coefficients. The matrix describes independent binding while the pair coefficients describe pair-wise interactions between peptide positions in the SMM algorithm [102]. Its performance is better than that of

matrix methods and the method has been implemented in a online prediction server called SMM[103].

The ANN method introduced in section 2.1, has been used for prediction of peptide binding to MHC [104, 105] and its prediction performance compares favorably relative to other available prediction methods [106]. ANN gradually outperforms other methods with increasing number of peptide/MHC binding data points [26]. Prediction system NetMHC [107] employed ANN as one of its prediction engine. Various methods have been adopted to improve the prediction performance of the ANN method. Evolutionary algorithms have been used for processing the training dataset for ANN [104]. Superior performance was achieved by combining several neural networks derived using different sequence-encoding schemes as compared to neural networks derived using only a single sequence-encoding scheme [108].

HMMs are a class of probabilistic models that are generally applicable to time series or linear sequences [109]. In bioinformatics, HMM methods have been widely used for motif-detection. They have also been effective for prediction of MHC binders [110].

Decision trees are rule based classification methods [111]. A decision tree with the BONSAI computational learning algorithm was constructed for identification of sequence motifs that affect T-cell activation, using a database of previously reported MHC bound peptides that had positive or negative T-cell responses [112]. Classification trees and bump-hunting methods have also been applied to prediction of MHC binding [113].

SVM belongs to the discriminative supervised machine learning methods. They are an implementation of the *structural risk minimization* principle, which minimizes the upper bound on the expected generalization error, as opposed to the *empirical risk minimization*

principle used in conventional neural networks which minimizes the error on the training data [114]. This difference gives SVM a greater ability to generalize, and decreases the possibility of over-fitting. They have been used for prediction of MHC class I binding peptides [115, 116]. The advantage of SVM is that it can be trained on relatively small peptide datasets. Prediction systems, SVMHC [115] and MHC-BPS [117] which predict peptide binders of flexible lengths for 18 HLA class I and 12 class II alleles, are based on SVM. SVRMHC adopted a quantitative support vector machine regression (SVR) approach for prediction of MHC binding peptides [118]. A novel "11-factor encoding" scheme taking into account similarities in the physicochemical properties of the amino acids constituting the input peptides has been adopted by SVRMHC and better performance has been achieved. Biosupport vector machine, which is modified from a conventional SVM by introducing a biobasis function so that the nonnumerical attributes of amino acids can be recognized without a feature extraction process, has been applied to T-cell epitope prediction [119].

Other computational methods, such as discriminate analysis [120], multivariate regression [121], and tree-based ensemble classifier [122], have been explored for prediction of MHC binding peptides too. Instead of identifying putative T-cell epitopes through the prediction of MHC binding peptides, a direct method for predicting CTL epitopes from an antigenic sequence has been proposed [123]. This method is based on matrix, SVM, and ANN methods trained and tested directly on T-cell epitopes and non-epitopes. The method also can discriminate between T-cell epitopes and MHC binders (non-epitopes). An online system, CTLpred, was built based on this method.

Table 2.2 A summary of publicly available computational systems related to T-cell epitope prediction.

System	Methods	URL	Reference	Prediction Function
MAPPP cleavage	Motif & Kenetic analysis	www.mpiib-berlin.mpg.de/MAPPP/cleavage.html	[124]	proteasomal cleavage
NetChop 3.0	ANN	www.cbs.dtu.dk/services/NetChop	[125]	
PAProC	Statistical method	www.paproc.de	[126]	
Pcleavage	SVM	www.imtech.res.in/raghava/pcleavage/	[127]	
SVMTAP	SVM	www-bs.informatik.uni-tuebingen.de/WAPP	[6]	TAP binding
TAPPred	SVM	www.imtech.res.in/raghava/tappred/	[128]	
BIMAS	Matrix	www-bimas.cit.nih.gov/molbio/hla_bind	[97]	MHC binding
CTLPred	Matrix, ANN, SVM	www.imtech.res.in/raghava/ctlpred/	[123]	CTL epitope
EPIPREDICT	Matrix	www.epipredict.de	[98]	HLA class II
MAPPP MHC	Matrix	www.mpiib-berlin.mpg.de/MAPPP/binding.html	[124]	HLA binding
MHCPred 2.0	QSAR	www.jenner.ac.uk/MHCPred	[87]	MHC/TAP binding
MHC-BPS	SVM	bidd.cz3.nus.edu.sg/mhc/	[117]	HLA binding
NetMHC	ANN	www.cbs.dtu.dk/services/NetMHC	[107]	HLA binding
PREDEP	Protein Threading	margalit.huji.ac.il/Teppred/mhc-bind/	[71]	MHC class I
HLABinding	Adaptive double threading	atom.research.microsoft.com/hlabinding/	[129]	HLA class I

PRED ^{NOD}	Matrix	antigen.i2r.a-star.edu.sg/Ag7/	[99]	H-2 ^{g7}
PptideCheck	Matrix	www.peptidecheck.org	[130]	HLA class I
ProPred	Matrix	www.imtech.res.in/raghava/propred	[131]	HLA-DR
RANKPEP	Matrix	bio.dfci.harvard.edu/RANKPEP/	[100, 101]	MHC & Proteasomal cleavage
SMM	SMM	zlab.bu.edu/SMM/	[103]	HLA-A2
SVMHC	SVM	www-bs.informatik.uni-tuebingen.de/Services/SVMHC	[115]	HLA-A, HLA-B
SVRMHC	SVR	svrmhc.umn.edu/SVRMHCdb/	[132]	MHC
SYFPEITHI	Motif	www.syfpeithi.de	[22]	MHC
MHC-PATHWAY	Matrix	www.mhc-pathway.net	[133]	MHC class I pathway
NetCTL	Matrix	www.cbs.dtu.dk/services/NetCTL	[134]	Twelve HLA class I supertypes, MHC class I pathway
PEPVAC	Matrix	bio.dfci.harvard.edu/PEPVAC/	[135]	Five HLA supertypes & Proteasomal cleavage

2.3.3 Amino acid encoding methods

In prediction using sequence-based approaches, multiple amino acid encoding methods have been applied. The simplest one is the conventional sparse (orthogonal) encoding in which each amino acid is encoded as a 20-digit binary string with a unique position set to “1” and other positions set to “0” [136, 137]. For example the first two amino acids, by alphabetic order, alanine (A) and cysteine (C) are encoded by 10000000000000000000 and 01000000000000000000, respectively, and the last amino acid tyrosine (Y) is encoded by 00000000000000000001. In this method, the chemical properties of amino acids are ignored and very specific information about the training peptides is inputted to a prediction algorithm.

Another encoding scheme is to use 5-bit binary strings to represent amino acids. The zeros and ones in a 5-bit string can be arranged in 32 possible ways. There are 20 representations left if the representations with no or all ones, and those with 1 or 4 ones are removed. This leaves just enough representations to code for the 20 amino acids [138].

In BLOSUM encoding method, amino acids are encoded as the BLOSUM (BLOCKS SUBstitution Matrix) score [139]. The BLOSUM matrix contains prior knowledge about similarity of amino acids so that it helps the prediction algorithms to generalize [108].

Brusic *et al.* used two amino acid encoding methods, named “Rep 6” and “Rep 9” as shown in Table 2.3. “Rep 6” assigned a 6-digit string to represent an amino acid. Each digit was either a scalar value representing one of the five features, hydrophobicity, volume, charge, aromatic side chain, and hydrogen bonds, or a correction digit. The scalar values for hydrophobicity and size were mapped to a scale of 0 to 10. The scalar values

for the rest three features were obtained from the Database of Amino Acid Indices [140].

“Rep 9” used a feature-based grouping of amino acids. Amino acid features such as: hydrophobicity, positive charge, negative charge, aromatic side chain, aliphatic side chain, small size, bulky size, and two correction bits for distinguishing similar amino acids were used to represent amino acids [141]. Little difference on prediction performance was observed from using “Rep 6” and “Rep 9” amino acid encoding methods and the sparse encoding method.

Table 2.3 Two amino acid encoding methods, Rep 6 and Rep 9.

Amino Acid	Rep 6	Rep 9
Alanine (A)	53500a	100001000
Cysteine (C)	45500a	000001001
Aspartate (D)	a5001a	001001000
Glutamate (E)	a60010	001000100
Phenylalanine (F)	185a0a	100100101
Glycine (G)	525000	000001000
Histidine (H)	565010	010100111
Isoleucine (I)	37500a	100010011
Lysine (K)	a7a02a	010000111
Leucine (L)	375000	100010001
Methionine (M)	375005	100000100
Asparagine (N)	655020	000001000
Proline (P)	555000	100000010
Glutamine (Q)	66502a	000000100
Arginine (R)	a8a046	010000110
Serine (S)	645010	000001011
Threonine (T)	55501a	000001010
Valine (V)	365007	100011010
Tryptophan (W)	0a5a10	100100110
Tyrosone (Y)	285a18	000100111
Any AA (X)	555555	111111111

Zhao *et al.* encoded an amino acid by ten factors [116]. These orthogonal factors were obtained from 188 physical properties of the 20 amino acids via multivariate statistical

analyses [142]. They account for 86% of the variance of the 188 physical properties. These factors included alpha-helix or bend-structure preference, bulk, beta-structure preference, hydrophobicity, normalized frequency of double bend, normalized frequency of alpha-region, and pK-C. This encoding method reduces the dimension of input vectors by half while enabling structural and biophysical properties to be better represented.

Milik *et al.* employed two amino acid encoding methods [143]. One is based on hierarchic organization of amino acid features formulated by Taylor [144]. Taylor clustered amino acids according to nine features, hydrophobic (H, W, Y, F, M, L, I, V, C, A, G, T, K), aliphatic (L, I, V), aromatic (F, Y, W, H), polar (T, S, N, D, E, Q, R, K, H, W, Y), charged (D, E, R, K, H), positive (R, K, H), small (P, V, V, A, G, T, S, N, D), tiny (A, G, S), proline (P). An additional class for Gly (G) was added due to the conformational flexibility of this amino acid. Thus vectors of 10 binary numbers (a “1” or “0” indicating the presence or absence of a given feature) were used to represent the 20 amino acids. Another encoding method is based on six physico-chemical properties of amino acids, volume, bulkiness, flexibility, polarity, aromaticity and charge.

2.3.4 Hybrid approaches for prediction of MHC binding peptides

Recently structure- and sequence-based methods were reported for MHC binding peptides prediction. One is named adaptive double threading, an improved protein threading approach which uses not only MHC/peptide complex structure data but also additional information from MHC sequences and binding affinities between MHC and peptides [129]. The built model significantly outperforms the standard threading approach in binding energy prediction because of the additional information used. In addition, to

combine both structural and sequence information for better prediction performance, the authors also built models for those MHC alleles without known binding peptides by using information from related, experimentally scrutinized alleles. The method performs especially well for the alleles with little or no experimental binding data available. HLABinding, a publicly available online prediction server, is the implementation of this method [129].

Another method, termed DynaPred, is a SVM-trained, quantitative matrix-based method with the features of the scoring matrix being energy terms retrieved from molecular dynamics simulations [145]. It is a two-step method. First, two feature matrices, one position-dependent and one position-independent, are constructed based on the linear binding free energy approximation of all 20 amino acids in each of the nine binding pockets of the MHC binding groove obtained from molecular dynamics simulation. Then the feature vectors act as input to a SVM model with a radial kernel function. The position-independent model showed low performance but higher generalization capability when compared to the position-dependent model. Its advantage is that it allows a fast prediction of MHC class I binding peptides and a construction of docked peptide conformation at the same time. However, the bound conformations of the residues in each binding pocket must be pre-calculated to facilitate a fast and efficient prediction process.

Combined computational methods that integrate multiple critical steps – proteasome cleavage, TAP transport, and MHC class I binding have been proposed as a supporting methodology for prediction of high probability targets for therapeutic peptides and vaccines [8]. Several combined computational applications of models of antigen processing and presentation have been reported, such as NetCTL and MHC-PATHWAY

[6, 103, 133, 134, 146]. A major weakness of these methods is that they represent a major antigen processing pathway, while alternative pathways also exist [7, 147] and hence the complete biological processes were not modeled. Of the steps involved, only peptide binding to MHC molecules is well-defined. Proteasome cleavage is a statistical event with preference for, but not exclusive to, specific cleavage sites. Processing involving other proteases also happens. Infected cells produce the immunoproteasome which has a completely different cleavage preference as compared to the constitutive proteasome. TAP binding is responsible for 90-97% of peptide processing depending on the HLA allele [136]. Signal peptides, membrane peptides, and some viral peptides are processed through TAP-independent pathways. Also, TAP can transport peptides in a form that is longer than the optimal HLA-binding length, with endoproteases trimming the peptide to the optimal size in the ER. Taken together, a simple sequential combination of prediction systems is not adequate. Both relevance and adequacy of such combined systems should be taken with caution.

The models for a small number of the most common human and mouse MHC alleles have been developed [148] but the small number of available experimental data for the majority of HLA alleles is still the main obstacle for predicting promiscuous T-cell epitopes by sequence-based methods. In this project, several sequence-based methods have been employed in building computational models for prediction of peptide binding to MHC molecules of both human and BALB/c mouse. The system for prediction of peptide binding to MHC molecules in human predicts peptides that bind to multiple HLA alleles belonging to one HLA supertype by a single prediction model per supertype. It can also predict peptide binding to HLA alleles for which no experimental binding data are

currently available. The details are reported in the following chapters.

2.4 Computational Methods Employed in Our Work

In our work, BP neural network, HMM, SVM, and quantitative matrix methods have been applied for prediction of peptides binding to MHC and TAP. We introduce these four methods in detail here.

2.4.1 Quantitative matrices

Quantitative matrices have been developed for prediction of peptide binding to H2^d molecules in Chapter 3. Let us consider a training dataset $T = (X, B)$ containing 9-mer peptides $X = \{X_j\}$ with known binding affinities $B = \{B_j\}$ to MHC molecules, where $j = \{1, 2, \dots, N\}$ and N is the number of peptides. The known binding affinity B_j is an integer ranging from 1 to 9. A peptide $X_j = \{x_{i,j}\}$ with binding affinity $B_j \geq 4$ is considered as a binder whereas with binding affinity $B_j < 4$ is considered as a non-binder, where $i = \{1, 2, \dots, 9\}$ denotes the amino acid position in a 9-mer peptide. $p_{a,i}$, the proportion of a particular amino acid a at the observed position i within binders, where $a \in \Omega$ and Ω denotes the 20 naturally-occurring amino acids, is defined as follows:

$$p_{a,i} = \frac{\sum_{j=1}^N \alpha \beta}{\sum_{j=1}^N \alpha} \quad \text{with } \alpha = \begin{cases} 1 & \text{if } B_j \geq 4 \\ 0 & \text{if } B_j < 4 \end{cases} \quad \text{and} \quad \beta = \begin{cases} 1 & \text{if } x_{i,j} = a \\ 0 & \text{if } x_{i,j} \neq a \end{cases} \quad (2.1)$$

The proportion of a particular amino acid a at the observed position i within non-binders, $n_{a,i}$, is defined as follows:

$$n_{a,i} = \frac{\sum_{j=1}^N \alpha \beta}{\sum_{j=1}^N \alpha} \quad \text{with } \alpha = \begin{cases} 1 & \text{if } B_j < 4 \\ 0 & \text{if } B_j \geq 4 \end{cases} \quad \text{and} \quad \beta = \begin{cases} 1 & \text{if } x_{i,j} = a \\ 0 & \text{if } x_{i,j} \neq a \end{cases} \quad (2.2)$$

The quantitative matrices $Q = \{q_{a,i}\}_{20 \times 9}$ were derived using logarithmic equations based on the frequency of amino acids at specific positions as described in [26].

$$q_{a,i} = \log_2 \left[\frac{p_{a,i} - n_{a,i}}{p_{a,i} + n_{a,i}} + 1 \right] \quad \text{if } p_{a,i} > n_{a,i} \quad (2.3)$$

$$q_{a,i} = -\log_2 \left[\frac{n_{a,i} - p_{a,i}}{p_{a,i} + n_{a,i}} + 1 \right] \quad \text{if } p_{a,i} < n_{a,i} \quad (2.4)$$

Where $q_{a,i}$ is the matrix coefficient for a particular amino acid at position i in the peptide.

The columns of the matrices were weighted using weight matrix, $W = \{w_i\}_{9 \times 1}$, according to the reported binding motifs. The matrix, Q , was refined by including information on the consensus [22] and other binding motifs; for example, most H2-K^d binding peptides contain Y, I, L, or M at anchor position P2 [94], and I, L, and V at anchor position P9. The anchor positions (e.g. P2 and P9 for K^d binding peptides) were assigned higher weights than other positions. The final models were tested and validated rigorously. After the matrix Q has been built, given a new 9-mer peptide $Y = \{y_i\}$, where $i = \{1, 2, \dots, 9\}$ denotes the amino acid position in a 9-mer peptide, with unknown binding affinity, its binding score S is defined as follows:

$$S = \sum_{i=1}^9 \sum_{a \in \Omega} q_{a,i} \cdot \delta \cdot w_i \quad \text{with } \delta = \begin{cases} 1 & \text{if } y_i = a \\ 0 & \text{if } y_i \neq a \end{cases} \quad (2.5)$$

2.4.2 Backpropagation neural networks

Backpropagation (BP) algorithm based on gradient descent in error is one of the most popular learning algorithms for neural networks [149]. The basic approach in network

learning is to start with an untrained network with random assigned weights, present a training pattern to the input layer, pass the signals through the net and determine the output at the output layer. The outputs are compared to the target outputs; any difference corresponds to an error. The weights are then adjusted to reduce the error.

Let us consider a training dataset $T = (X, B)$ containing 9-mer peptides $X = \{X_j\}$ with known binding affinities $B = \{B_j\}$ to MHC molecules, where $j = \{1, 2, \dots, N\}$ and N is the number of peptides. The inputs to the ANN were the binary strings representing 9-mer peptides. The outputs of the ANN models were binding scores.

Figure 2.1 A three layer feed-forward network.

Let us consider a three-layer network as illustrated in Figure 2.1. The network is a set of neurons that are connected to each other via edges. Input neurons are denoted by I_k ,

hidden neurons are denoted by H_m , and the output neuron is denoted by O . In our work, the number of output neurons is one, as the output is a value representing the predicted binding score. A neuron sums up all its received "signals" and inputs it into an activation function, then sends the result to all its children. This signal is multiplied by a weight associated with an edge, then is received by the target neuron. The weights associated with the edges from input neurons to hidden neurons and from hidden neurons to the output neuron are denoted by W_{km} and G_m , respectively. Let the number of input neurons be K and the number of hidden neurons be M . The architecture of the neural network shown in Figure 2.1 is K - M -1. Give an input peptide X_j , the input to a hidden neuron H_m is

$$H_m^j = \sum_{k=1}^K W_{km} I_k^j \quad (2.6)$$

After applying activation function f , the resulted output of the hidden neuron is

$$V_m^j = f(H_m^j) = f\left(\sum_{k=1}^K W_{km} I_k^j\right) \quad (2.7)$$

The input to the output neuron is

$$L^j = \sum_{m=1}^M G_m V_m^j = \sum_{m=1}^M G_m f\left(\sum_{k=1}^K W_{km} I_k^j\right) \quad (2.8)$$

After going through activation function g , the output is

$$O^j = g(L^j) = g\left(\sum_{m=1}^M G_m V_m^j\right) = g\left(\sum_{m=1}^M G_m f\left(\sum_{k=1}^K W_{km} I_k^j\right)\right) \quad (2.9)$$

The output error is

$$\begin{aligned} E &= \frac{1}{2N} \sum_{j=1}^N [B_j - O^j]^2 \\ &= \frac{1}{2N} \sum_{j=1}^N [B_j - g\left(\sum_{m=1}^M G_m f\left(\sum_{k=1}^K W_{km} I_k^j\right)\right)]^2 \end{aligned} \quad (2.10)$$

Where B_j is the expected output for a peptide X_j .

For the hidden-to-output connections the gradient descent rule gives

$$\begin{aligned}\Delta G_m &= -\eta \frac{\partial E}{\partial G_m} = \frac{\eta}{N} \sum_{j=1}^N [B_j - O^j] \bullet g'(L^j) V_m^j \\ &= \frac{\eta}{N} \sum_{j=1}^N \delta^j V_m^j\end{aligned}\quad (2.11)$$

Where $\delta^j = g'(L^j)[B_j - O^j]$ and η is learning rate.

The selection of a learning rate is important in finding the global minimum of the error distance. Too small a learning rate makes the network take a long time to converge. Too large a learning rate makes the network converge fast, but the network may end up finding a local minimum of the error distance.

For the input-to-hidden connections ΔW_{km} ,

$$\begin{aligned}\Delta W_{km} &= -\eta \frac{\partial E}{\partial W_{km}} = -\eta \frac{\partial E}{\partial V_m^j} \frac{\partial V_m^j}{\partial W_{km}} \\ &= \frac{\eta}{N} \sum_{j=1}^N [B_j - O^j] g'(L^j) G_m f'(H_m^j) I_k^j \\ &= \frac{\eta}{N} \sum_{j=1}^N \delta^j G_m f'(H_m^j) I_k^j \\ &= \frac{\eta}{N} \sum_{j=1}^N \theta^j I_k^j\end{aligned}\quad (2.12)$$

where $\theta^j = f'(H_m^j) G_m \delta^j$.

Empirical evidence shows that the use of momentum in the BP algorithm helps in speeding the convergence and avoiding local minima. After applying momentum ε , ΔG_m as shown in equation (2.11) and ΔW_{km} as shown in equation (2.12) now change to:

$$\begin{aligned}\Delta G_m(t) &= -\eta \frac{\partial E}{\partial G_m} + \varepsilon \Delta G_m(t-1) \\ &= \frac{\eta}{N} \sum_{j=1}^N \delta^j V_m^j + \varepsilon \Delta G_m(t-1)\end{aligned}\quad (2.13)$$

$$\begin{aligned}
\Delta W_{km}(t) &= -\eta \frac{\partial E}{\partial W_{km}} + \varepsilon \Delta W_{km}(t-1) \\
&= \frac{\eta}{N} \sum_{j=1}^N \theta^j I_k^j + \varepsilon \Delta W_{km}(t-1)
\end{aligned} \tag{2.14}$$

Where t denotes the number training epochs. Using a momentum stabilizes the weight change by using a combination of the gradient decreasing term with a fraction of the previous weight change. It also enables the escape from small local minima and amplifies the learning rate causing a faster convergence when all weight changes are all in the same direction.

The steps involved in training a BP neural network are as below:

-
- 1: **Input:** $T = (X, B)$, $X = \{X_j\}$, $B = \{B_j\}$, $j = \{1, 2, \dots, N\}$
 - 2: **Output:** W_{km} and G_m
 - 3: BEGIN
 - 4: Initialize weights W_{km} and G_m to random values
 - 5: **Repeat**
 - 6: **for** each peptide X_j in the training dataset **do**
 - 7: Calculate O^j using equation (2.6) to (2.9)
 - 8: **end for**
 - 9: Calculate ΔG_m and ΔW_{km} using equation (2.11) to (2.12)
 - 10: Update weights $G_m = G_m + \Delta G_m$ and $W_{km} = W_{km} + \Delta W_{km}$
 - 11: **Until** stopping criterion is satisfied
 - 12: **Return** W_{km} and G_m
 - 13: END
-

Frequently used activation functions include linear, sigmoid and hyperbolic tangent functions. A sigmoid is an S-shaped curve that maps all input to $[0, 1]$:

$$\text{sgm}(x) = \frac{1}{1 + e^{-x}} \quad (2.15)$$

Hyperbolic tangent function maps all input to $[-1, 1]$:

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (2.16)$$

This backpropagation algorithm has been employed in developing ANN models for prediction of peptides binding to TAP and MHC in Chapter 4 and 5.

2.4.3 Hidden Markov models

The basic theory of Hidden Markov models was published in late 1960s and early 1970s [109, 150]. It has been widely applied in speech recognition field since the 1970s [151] and was introduced into computational biology in the late 1980s [152]. It has been proven successful for modeling biological sequences [153] and prediction of biological sequence structure [154].

Let us consider a system that can be described as being in one of the states in a set of N distinct states, $(s_1, s_2, \dots, s_N)$, as illustrated in Figure 2.2 (where $N = 5$, for simplicity). Time associated with state changes is denoted as $t = 1, 2, \dots$, and the state at time t is denoted as q_t . For a discrete, first order Markov chain, the probabilistic description can be written as

$$\begin{aligned} P[q_t = s_i \mid q_{t-1} = s_j, q_{t-2} = s_k, \dots] \\ = P[q_t = s_i \mid q_{t-1} = s_j] \\ 1 \leq i, j, k \leq N \end{aligned} \quad (2.17)$$

Figure 2.2 A Markov chain with five states.

If the right-hand side of equation (2.17) is independent of time, it leads to the set of state transition probabilities $p_{i,j}$ of the form

$$p_{i,j} = P[q_t = s_i | q_{t-1} = s_j] \quad 1 \leq i, j \leq N \quad (2.18)$$

with $p_{i,j} \geq 0$ and $\sum_{j=1}^N p_{i,j} = 1$ as they need to obey standard probability constraints. This

is an observable Markov model since the output of the process is the set of states at each instant of time, where each state corresponds to a physical (observable) event. There are a number of extensions to classical Markov chains and our focus is HMM. An HMM can be characterized as following:

- 1) N , the number of states in the model. The individual states are denoted as $S = \{s_1, s_2, \dots, s_N\}$, and the state at time t as q_t . In this work, N is 20 and $\{s_1, s_2, \dots, s_N\}$ are the 20 amino acids.
- 2) M , the number of distinct observation symbols per state, i.e., the discrete alphabet size. The individual symbols are represented as $V = \{v_1, v_2, \dots, v_M\}$.

- 3) The state transition probability distribution $P = \{p_{i,j}\}$, where

$$p_{i,j} = P[q_t = s_i \mid q_{t-1} = s_j] \quad 1 \leq i, j \leq N$$

- 4) The observation symbol probability distribution in state j , $B = \{b_j(k)\}$, where

$$b_j(k) = P[v_k \text{ at } t \mid q_t = s_j] \quad 1 \leq j \leq N, 1 \leq k \leq M$$

- 5) The initial state distribution, $W = \{w_i\}$, where

$$w_i = P[q_1 = s_i] \quad 1 \leq i \leq N$$

Given appropriate values of N , M , P , B and W , the HMM can be used to generate an observation sequence $O = O_1, O_2, \dots, O_T$, where T is the number of observations.

There are three basic problems of HMMs,

- 1) Given the observation sequences $O = O_1, O_2, \dots, O_T$, and a model $\lambda = \{P, B, W\}$, how to compute the probability of the observation sequences, $p(O \mid \lambda)$? This is also called the evaluation or scoring problem.
- 2) Given the observation sequences $O = O_1, O_2, \dots, O_T$, and a model $\lambda = \{P, B, W\}$, how to choose a corresponding state sequence $O = O_1, O_2, \dots, O_T$, which is optimal in some sense? This problem is called the alignment or decoding problem.
- 3) How to adjust the model parameters $\lambda = \{P, B, W\}$ to maximize $p(O \mid \lambda)$? This is called the training or estimation problem.

In problem 3, we try to optimize the model parameters to best describe the observation sequence. The observation sequence is called a training sequence since it is used to train the HMM. By using an iterative procedure such as the Baum-Welch method or EM (expectation-modification) method, or gradient techniques, $\lambda = \{P, B, W\}$ can be chosen so that $p(O \mid \lambda)$ is locally maximized.

Let the probability of the partial observation sequence, $O_1, O_2, \dots, O_t$, and state s_i at

time t , given the model λ be $\alpha_t(i) = P(O_1 O_2 \cdots O_t, q_t = s_i | \lambda)$. Let the probability of the partial observation sequence, $O_{t+1}, \dots, O_T$, given state s_i at time t and the model λ be $\beta_t(i) = P(O_{t+1} O_{t+2} \cdots O_T | q_t = s_i, \lambda)$. The detailed calculation of $\alpha_t(i)$ and $\beta_t(i)$ can be found in [155]. Let $\xi_t(i, j)$ be the probability of being in state s_i at time t and state s_j at time $t+1$. Then $\xi_t(i, j) = P(q_t = s_i, q_{t+1} = s_j | O, \lambda)$.

$$\xi_t(i, j) = \frac{\alpha_t(i) p_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{P(O | \lambda)} = \frac{\alpha_t(i) p_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^N \sum_{j=1}^N \alpha_t(i) p_{ij} b_j(O_{t+1}) \beta_{t+1}(j)} \quad (2.19)$$

We define $\gamma_t(i)$ as the probability of being in state s_i at time t given the observation sequence and the model λ . $\gamma_t(i)$ can be related to $\xi_t(i, j)$ by summing over j , giving $\gamma_t(i) = \sum_{j=1}^N \xi_t(i, j)$. $\sum_{t=1}^{T-1} \gamma_t(i)$ can be interpreted as the expected (over time) number of times that state S_i is visited. $\sum_{t=1}^T \xi_t(i, j)$ can be interpreted as the expected (over time) number of transitions from state s_i to s_j .

Given the above formulas, the parameter estimation formulas for W , P and B are

$$\overline{W}_j = \text{expected frequency in state } s_i \text{ at time } (t=1) = \gamma_1(i) \quad (2.20)$$

$$\begin{aligned} \overline{p}_{ij} &= \frac{\text{expected number of transitions from state } s_i \text{ to } s_j}{\text{expected number of transitions from state } s_i} \\ &= \frac{\sum_{t=1}^{T-1} \xi_t(i, j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \end{aligned} \quad (2.21)$$

$$\begin{aligned} \overline{b}_j(k) &= \frac{\text{expected number of times in state } s_j \text{ and observing symbol } v_k}{\text{expected number of times in state } s_j} \\ &= \frac{\sum_{t=1}^T \gamma_t(j)}{\sum_{t=1}^T \gamma_t(j)} \end{aligned} \quad (2.22)$$

If we use the current model $\lambda = \{P, B, W\}$ to compute the right-hand side of equation

(2.20) - (2.22), and we define the re-estimated model as $\bar{\lambda} = \{\bar{P}, \bar{B}, \bar{W}\}$ as computed from the left-hand sides of (2.20) - (2.22). If $\bar{\lambda}$ is iteratively used in place of λ in the re-estimation calculations, then the probability of O being observed from the model can be improved until some limiting point is reached. The final result of this procedure is called a maximum likelihood estimate of the HMM.

A Dirichlet mixture can be used to improve the detection of weak but significant protein sequence homology. Dirichlet mixture densities have been combined with observed amino acid frequencies to form estimates of expected amino acid probabilities at each position in a profile or hidden Markov model [156]. A 9-component Dirichlet mixture has been used to give priors over amino acid distributions in our HMM models.

During evolution, the probabilities of different amino acid substitutions are very different. Taking this into account can greatly improve the performance of the HMM models built. BLOSUM took into consideration evolutionary distances by ignoring the sequences that had more than a certain percent identity [139]. For example, the substitution scores of BLOSUM62 matrix were derived from the alignments of sequences that had no more than 62% identity. In our work, BLOSUM62 has been used to weight the input peptide sequences.

In the present study, we used a first-order HMM, in which the current system state is determined only by the preceding state. The process used for building the HMM [157] included the following steps:

- 1) Input of peptides;
- 2) Calculate distances between the peptides to cluster them;
- 3) Weight each input peptide using the BLOSUM62 matrix and scaling;

- 4) Use Dirichlet mixtures to estimate prior probability distributions associated with the model; and model configuration.
- 5) Construct HMM.

The Viterbi algorithm [158] is a well known way for computing the best-scoring path through an HMM by scoring, in parallel, each path through the HMM and keeping track of the best ones. When the HMM models were build, given a new peptide sequence, the Viterbi algorithm was used for calculate its binding score.

In a first order HMM, there are a few assumptions: the Markov assumption which assumes that the current state is only dependent on the previous state; the stationary assumption which assumes that the transition probabilities are independent of the actual time at which the transitions take place; the output independence assumption which states that the current observation is statistically independent of the previous observations (<http://jedlik.phy.bme.hu/~gerjanos/HMM/node5.html>). In the case of MHC binding peptide prediction, the last two assumptions are valid but the first assumption is not valid as the interactions among amino acids in a peptide may not be limited only to their direct neighbors in front of them. This invalid assumption may affect the prediction performance of the HMMs built.

2.4.4 Support vector machines

SVMs are becoming very popular due to many attractive features and promising empirical performance. SVMs implement a simple idea: they map pattern vectors to a high-dimensional feature space where a ‘best’ separating hyper-plane is constructed (Webb, 2002). It provides a linear separation in an augmented space, different from the

original one, by means of some defined kernels. SVM are an implementation of the *structural risk minimization* principle [114], which minimizes the upper bound on the expected generalization error, opposed to the *empirical risk minimization* principle, used in conventional neural networks, which minimizes the error on the training data. It is this difference which equips SVM with a greater ability to generalize, which is the goal in statistical learning [159].

Here we consider the binary classification task in which we have a set of training patterns $\{x_i\}$, where $i=1,\dots,n$, assigned to one of two classes, ω_1 and ω_2 , with corresponding labels $g(x)=0$. In this study, x_i is a binary string representing a 9-mer peptide and class ω_1 and ω_2 correspond to HLA binders and non-binders. Denote the linear discriminate function

$$g(x_i) = w^T x_i + w_0 \quad (2.23)$$

with decision rule

$$w^T x + w_0 \begin{cases} > 0 \\ < 0 \end{cases} \Rightarrow x \in \begin{cases} \omega_1 \text{ with corresponding numeric value } y_i = +1 \\ \omega_2 \text{ with corresponding numeric value } y_i = -1 \end{cases} \quad (2.24)$$

Thus, all training points are correctly classified if

$$y_i(w^T x_i + w_0) > 0 \quad \text{for all } i$$

There may be more than one separating hyperplane. The maximal margin classifier determines the hyperplane for which the margin – the distance to two parallel hyperplanes on each side of the separating hyperplane – is the largest. The assumption is that the larger the margin, the better the generalization error of the linear classifier defined by the separating hyperplane. To aid generalization, a margin, $b > 0$, is introduced. The problem now is to seek a solution so that

$$y_i(w^T x_i + w_0) \geq b \quad (2.25)$$

Without loss of generality, a value $b=1$ may be taken, defining what are termed the *canonical hyperplanes*, $H_1: w^T x_i + w_0 = +1$ and $H_2: w^T x_i + w_0 = -1$ and we have

$$w^T x + w_0 \begin{cases} \geq +1 \\ \leq -1 \end{cases} \Rightarrow x \in \begin{cases} \omega_1 \text{ with corresponding numeric value } y_i = +1 \\ \omega_2 \text{ with corresponding numeric value } y_i = -1 \end{cases} \quad (2.26)$$

The distance between each of these two hyperplanes and the separating hyperplane, $g(x)=0$, is $1/|w|$ and is termed the *margin*. Maximizing the margin means that we seek a solution that minimizes $|w|$ subject to the constraints

$$y_i(w^T x_i + w_0) \geq 1 \quad i = 1, \dots, n \quad (2.27)$$

A standard approach to optimization problems with equality and inequality constraints is the Lagrange formalism [160] which leads to the primal form of the objective function, L_p , given by

$$L_p = \frac{1}{2} w^T w - \sum_{i=1}^n \alpha_i (y_i (w^T x_i + w_0) - 1) \quad (2.28)$$

where $\{\alpha_i, i = 1, \dots, n; \alpha_i \geq 0\}$ are the Lagrange multipliers. The solution to the problem of minimizing $w^T w$ subject to constraints (2.27) is equivalent to determining the saddle point of the function L_p , at which L_p is minimized with respect to w and w_0 and maximized with respect to the α_i . Differentiating L_p with respect to w and w_0 and equating to zero yields

$$\sum_{i=1}^n \alpha_i y_i = 0 \quad (2.29)$$

$$w = \sum_{i=1}^n \alpha_i y_i x_i$$

Substituting into equation (2.28) gives the dual form of the Lagrangian

$$L_D = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j x_i^T x_j \quad (2.30)$$

which is maximized with respect to the α_i subject to

$$\alpha_i \geq 0 \quad \sum_{i=1}^n \alpha_i y_i = 0 \quad (2.31)$$

The support vector algorithm may be applied in a transformed feature space, $\phi(x)$, using some nonlinear function. The discriminate function as shown in equation (2.23) now changes to

$$g(x) = (w^T \phi(x) + w_x) \quad (2.32)$$

Acceptable kernels must be expressible as an inner product in a feature space, or $k(x, y) = \phi^T(x)\phi(y)$. Frequently used kernels are polynomial

$$k(x, y) = (1 + x^T y)^d \quad (2.33)$$

and radial basis (Gaussian) kernel

$$k(x, y) = \exp\left(-g\|x - y\|^2\right) \quad (2.34)$$

However, a separating hyperplane does not always exist, even in the feature space, e.g. when there is a high level of noise that causes an overlap of the classes. The standard approach is to allow the possibility of some misclassified examples by introducing a *soft margin* optimal hyperplane. We again have the same quadratic programming problem, but subject to one different constraint:

$$0 \leq \alpha_i \leq C \quad (2.35)$$

The upper bound C controls the influence of misclassified examples, and represents a trade-off between training error and margin.

Beside classification problems, SVM can also be applied to regression problems by introducing a loss function which includes a distance measure [161].

2.5 Validation Methods and Performance Measures Used in Our Work

2.5.1 Cross validation

Cross validation is a method for error rate estimation [162]. It implements a simple idea: the dataset of size n samples is partitioned into two parts, the model parameters are estimated using one set and the goodness-of-fit criterion evaluated on the second set. The cross-validation estimates the goodness-of-fit criterion. Cross validation tends to over fit when selecting a correct model – it may choose an overly-complex model for the given dataset. There is some evidence that for model selection, multifold cross validation, where more than one sample is deleted from the training set in each comparison, performs better than a simple leave-one-out cross validation [163]. In most of our experiments, 10-fold cross validation was performed to evaluate the performance of the classifiers.

2.5.2 t-test

The t-test assesses whether the means of two groups are statistically different from each other [164, 165]. Statistical significance is the degree of risk one is willing to take that one rejects a null hypothesis when it is actually true. In our case, the null hypothesis says there is no difference between the two sets of predictions made by two different methods.

The formula for computing the t value for independent means is shown as below,

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left[\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \right] \left[\frac{n_1 + n_2}{n_1 n_2} \right]}}$$

Where $\bar{X}_1$ is the mean of data group 1, $\bar{X}_2$ is the mean of data group 2, n_1 is the number of data points in group 1, n_2 is the number of data points in group 2, s_1^2 is the

variance for group 1 and s_2^2 is the variance for group 2. The formula for computing the t value for dependent means is shown as below,

$$t = \frac{\sum D}{\sqrt{\frac{n \sum D^2 - (\sum D)^2}{n-1}}}$$

Where $\sum D$ is the sum of all the differences between groups and n is the number of pairs of observations. In Chapter 5, a t-test was performed to compare the performance of several classifiers.

2.5.3 Performance measures

In our work, the predictive performance of the models was evaluated by sensitivity (SE) and specificity (SP) measures. Sensitivity, $SE = TP / (TP + FN)$, indicates the percentage of correctly predicted binders, where TP stands for number of true positive predictions (experimental binders predicted as binders) and FN stands for number of false negative predictions (experimental binders predicted as non-binders). Specificity, $SP = TN / (TN + FP)$, indicates the percentage of correctly predicted non-binders, where TN stands for number of true negative predictions (experimental non-binders predicted as non-binders) and FP stands for number of false positive predictions (experimental non-binders predicted as binders). There is a trade-off between specificity and sensitivity: the higher the specificity, the lower the sensitivity, which results in lower values for both TP and FP. The lower the specificity, the higher the sensitivity, which results in a higher number of TP at cost of a higher number of FP. We consider values of $SP \geq 0.8$ useful in practice and assessed SE for three values of SP (0.8, 0.9 and 0.95).

The receiver operating characteristic (ROC) curve analysis provided the overall

prediction accuracies [166]. The ROC curve is a plot of the SE vs. (1 - SP) for a binary classifier as its discrimination threshold is varied. As a rough guide, the area under ROC (AROC) value varies from 0 to 1.0, with 1.0 representing a perfect prediction, 0.9 to 1.0 representing excellent accuracy, 0.8 to 0.9 representing good accuracy, 0.7 to 0.8 representing minimum acceptable accuracy, and 0.5 to 0.7 representing predictions that indicate random choice.

To assess the accuracy of binding affinity predictions we calculated the Pearson correlation coefficient between the known binding scores X of a set of peptides and their prediction scores Y.

$$r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

x_i and $\bar{x}$ are known binding score and average binding scores; y_i and $\bar{y}$ are individual and average peptide predictions. Pearson correlation coefficient, ranging from +1 to -1, reflects the degree of linear relationship between two variables. A correlation of +1 represents a perfect positive linear relationship between variables, while a correlation of -1 means a perfect negative linear relationship between variables. A correlation of 0 means no linear relationship between the two variables [167].

Chapter 3 **PRED^{BALB/c}: a system for prediction of peptide binding to the H2^d molecules, a haplotype of the BALB/c mouse**

The BALB/c inbred laboratory mouse strain is one of the most commonly used animal models in immunological studies and has been extensively used in vaccine research [168, 169]. BALB/c mice express three MHC class I (H2-K^d, -L^d and -D^d) and two class II (I-A^d and -E^d) molecules.

3.1 Introduction

Several publicly available prediction systems for MHC class I and class II binding peptides provide the prediction models for H2^d alleles. SYFPEITHI [22] has H2-K^d and -L^d models, BIMAS [97] has H2-K^d, -D^d, and -L^d models, libscore [170] has H2-L^d model, and RANKPEP [100] has models for all H2^d molecules. SYFPEITHI uses binding motifs, BIMAS and RANKPEP use binding matrices, and libscore carries out positional scanning using combinatorial peptide libraries to analyze MHC binding peptides. These are general servers that contain prediction models for a range of MHC molecules in human and mouse, and several other mammalian organisms, but the accuracies of most individual models have not been determined. On the other hand, quantitative matrices for H2-K^b, -D^b, -L^d, and -K^k [170, 171] have been developed and validated.

We have developed a computational system, PRED^{BALB/c}, that predicts peptide binding to five H2^d alleles of the BALB/c mouse. The primary goal was to develop a computational model for accurate prediction of peptide binding to all five alleles of H2^d molecules and integrate them for accurate prediction of immunodominant T-cell epitope

regions in the context of the whole organism. The prediction system utilizes quantitative matrices, which were rigorously validated using experimentally determined binders and non-binders and also by in vivo studies using viral proteins. PRED^{BALB/c} is the first online server for prediction of peptide binding to a complete set of MHC molecules in a model organism (*i.e.* H2^d haplotype). PRED^{BALB/c} is accessible at <http://antigen.i2r.a-star.edu.sg/predBalbc/>. The user instructions for PRED^{BALB/c} are in the Appendix A.

3.2 Material and Methods

3.2.1 Training datasets and data preprocessing

The training datasets containing both binding and non-binding peptides were extracted from MHCPEP [21], MHCBN [172], SYFPEITHI [22], and JenPep [173]. The 9-mer peptides were used for deriving H2^d class I matrices because majority of peptides that bind these molecules are nine amino acid long [174]. Although the majority of H2^d class II binding peptides are 12-20 amino acid long, their binding cores are nine amino acid long [62, 175].

The numbers of binders and non-binders in the training data are given in Table 3.1. The training data are available for download at <http://antigen.i2r.a-star.edu.sg/predBalbc/data>. In each data file, there are two columns: the first column contains 9-mer peptides and the second column contains the corresponding binding affinities. The binding affinities are denoted by four letters “N”, “H”, “M”, and “L”, representing non-binders, high binders, moderate binders, and low binders, respectively.

Table 3.1 Number of 9-mer peptides in the training datasets of PRED^{BALB/c} class I and class II models.

	Class I			Class II	
	H2-D ^d	H2-K ^d	H2-L ^d	I-A ^d	I-E ^d
Binders	93	139	73	127	191
Non binders	331	105	91	963	627

3.2.2 Prediction engine

Because relatively small training datasets are available for H2^d, we adopted quantitative matrices as the prediction method. Machine learning methods, such as ANN and SVM, outperform the matrix method as increasing amounts of binding data become available. The details on building quantitative matrices are available at section 2.4.1.

3.2.3 Normalization of prediction scores

The prediction scores calculated by matrix methods were further rescaled to map them into the range from one to nine. The mapping of scores was done using equation (3.1), where S_N denotes the normalized score, S denotes the original prediction score calculated using equation (2.5), $S_{\min}$ and $S_{\max}$ denote the possible minimum and maximum values of the original scores, respectively. The values for $S_{\min}$ and $S_{\max}$ were obtained through extensive testing. More than 650 randomly selected protein sequences from the NCBI protein database (contained more than 220,000 9-amino acid peptides) were used for prediction using the matrix models. Since the testing data contained a large number of 9-amino acid peptides, the highest and lowest predicted score from the testing data were taken as reasonable maximum and minimum scores for normalization.

$$S_N = (S - S_{\min}) / (S_{\max} - S_{\min}) * 8 + 1 \quad (3.1)$$

S_N was further mapped to a linear scale ranging from 0-10 that corresponds to specificity values of the prediction models so that the prediction thresholds across

different models would be identical. The specificity values at different thresholds (based on S_N) were obtained through 10-fold cross validation. For example, when a user sets the threshold to 8, the specificity of the predictions to all five alleles is 0.8. Raw binding scores are mapped to a linear scale that corresponds to specificity values and therefore the prediction thresholds across different models have similar meaning.

3.3 Results

An assessment of predictive accuracy was carried out for three subsets of peptide binders:

1) all binders including low, moderate, and high binders were considered as positive samples, and all non-binders as negative samples (referred to as the LMH set); 2) moderate and high binders were considered as positive samples, all non-binder and low binders as negative samples (referred to as the MH set), and 3) only high binders were considered as positive samples, with all other peptides as negative samples (referred to as the H set).

The performance of the $\text{PRED}^{\text{BALB/c}}$ models was evaluated by 10-fold cross validations, by a set of experimentally identified E^d and A^d binding data [176], and by some T-cell epitope regions identified by $\text{IFN-}\gamma$ ELISpot of the splenocytes of BALB/c immunized mice. The prediction performance of $\text{PRED}^{\text{BALB/c}}$ was compared to those of several publicly available predictive systems. The details on cross validation are available in section 2.5.1. The performance measurements used were introduced in section 2.5.3.

3.3.1 Ten-fold cross validation

Ten-fold cross validations were performed to test the accuracy of the prediction models.

The results show that $\text{PRED}^{\text{BALB/c}}$ predicts peptides binding to E^{d} , A^{d} and K^{d} with excellent accuracy ($\text{AROC} \geq 0.90$), and to D^{d} and L^{d} with good accuracy ($\text{AROC} \geq 0.85$).

Figure 3.1 shows the specificity and sensitivity relationship of the five prediction models.

A)

B)

55

C)

D)

56

E)

Figure 3.1 Sensitivities vs. specificities of the five PRED^{BALB/c} models obtained from 10-fold cross validation. (A) D^d, (B) K^d, (C) L^d, (D) A^d, and (E) E^d.

3.3.2 Validation of PRED^{BALB/c} models using independent data

H2^d class II models were also validated using a set of experimentally identified E^d and A^d binding data from several protein antigens [176]. To ensure independency of the training and the testing datasets, we compared the training and testing data and any overlapping 9-mer peptide was removed from training datasets and new prediction models were built based on the new training datasets. Peptides described by [176] were mostly longer than 9 amino acids and hence were decomposed into overlapping 9-mer peptide sequences with an 8 amino acid overlap. Each of the 9 amino acid peptides was submitted to E^d and A^d models of PRED^{BALB/c}. For non-binders, the prediction of all 9-mer peptides was taken, while for binders, only the high scoring 9-mer peptide of the parent peptide was considered and consequently, the performance of both models was thus assessed.

Sensitivity, specificity and AROC of the predictions were calculated; both A^d and E^d models performed with excellent accuracy ($AROC \geq 0.95$) (Table 3.2). Figure 3.2 shows the specificity and sensitivity relationship of the two prediction models.

Table 3.2 Sensitivities for $PRED^{BALB/c}$ (A) A^d , (B) E^d models for the set of synthetic peptides from eleven proteins with experimentally identified E^d and A^d binding. The values are given for three levels of specificity (0.8, 0.9, and 0.95) for prediction of high (H), moderate (M) and low (L) binders.

A)

Sensitivities for the prediction of I- A^d binding peptides			
Level	LMH	MH	H
SP = 0.80	0.97	0.97	1.00
SP = 0.90	0.83	0.83	0.89
SP = 0.95	0.62	0.62	0.42

B)

Sensitivities for the prediction of I- E^d binding peptides			
Level	LMH	MH	H
SP = 0.80	1.00	1.00	1.00
SP = 0.90	0.87	0.87	0.90
SP = 0.95	0.87	0.87	0.90

A)

Sensitivity vs. Specificity for I- A^d predictions

B)

Figure 3.2 Sensitivities vs. Specificity for PRED^{BALB/c} (A) I-A^d, (B) I-E^d models for the set of synthetic peptides from eleven proteins with experimentally identified E^d and A^d binding.

3.3.3 Experimental validation of PRED^{BALB/c} models

The severe acute respiratory syndrome coronavirus nucleocapsid (SARS-CoV N) T-cell epitope sequences identified by IFN- γ ELISpot of the splenocytes of BALB/c immunized mice were analyzed for the validation of the PRED^{BALB/c} system. BALB/c T-cell epitope predictions with SARS-CoV N protein show strong correlation to the experimental ELISpot data [177]. Peptides that elicited stimulatory effects in splenocytes from immunized mice with the SARS-CoV N vaccines were among those with the highest probability of predicted T-cell epitope sequences within the top 5% of predictions. The predicted 9-mer binders in SARS-CoV N by PRED^{BALB/c} included all of the nine IFN- γ ELISpot positive peptide sequences, N76-114 (5 overlapping peptides), N130-149, N241-258, N352-366 and N353-370 (Table 3.3A). The IFN- γ ELISpot positive region,

N76-114, covered 16 overlapping 9-mers predicted to bind to all five H2^d alleles by PRED^{BALB/c} (Table 3.3B). The IFN- γ ELISpot positive region, N130-149, covered five overlapping 9-mers predicted to bind to four H2^d alleles (Table 3.3C). The IFN- γ ELISpot positive region, N241-258, covered four overlapping 9-mers predicted to bind to H2^d class II allele A^d (Table 3.3D). The IFN- γ ELISpot positive region, N352-370, covered three 9-mers predicted to bind to three H2^d alleles (Table 3.3E).

Table 3.3 The top 5% of PRED^{BALB/c} predictions vs. T-cell IFN- γ ELISpot peptides in SARS-CoV N. PRED^{BALB/c} predictions covered all of the nine T-cell IFN- γ immunodominant ELISpot positive sequences from SARS-CoV N. (A) The T-cell IFN- γ ELISpot peptides of SARS-CoV N are indicated by the position number, sequence and PRED^{BALB/c} alleles predicted. (B) The IFN- γ ELISpot peptides spanning the region N76-114 contained 16 9-mers within the top 5% of predictions of all the five H2^d alleles. (C) The IFN- γ ELISpot peptides spanning the region N130-149 contained five 9-mer peptides that were within the top 5% of predictions of four H2^d alleles. (D) The IFN- γ ELISpot peptides spanning the region N241-258 contained four 9-mer peptides that were within the top 5% of predictions of A^d allele. (E) The IFN- γ ELISpot peptides spanning the region N352-370 contained three 9-mer peptides that were within the top 5% of predictions of three H2^d alleles.

A)

Position	SARS-CoV N ELISpot positive peptide sequences	PRED ^{BALB/c} allele predicted
76-93	NTNSGPDDQIGYYRRATR	D ^d
80-99	GPDDQIGYYRRATRRVRGGD	L ^d , K ^d , D ^d , A ^d , E ^d
84-101	QIGYYRRATRRVRGGDGK	K ^d , D ^d , A ^d , E ^d
92-109	TRRVRGGDGKMKELSPRW	D ^d , E ^d
100-114	GKMKELSPRWYFYLYL	K ^d , L ^d , D ^d , E ^d
130-149	GIVWVATEGALNTPKDHIGT	K ^d , L ^d , A ^d , E ^d
241-258	QQQGQTVTKKSAAEASKK	A ^d
352-366	ILLNKHIDAYKTFPP	K ^d , A ^d , E ^d
353-370	LLNKHIDAYKTFPPTEPK	K ^d , A ^d , E ^d

B)

PRED ^{BALB/c} H2 ^d alleles	N76-114 T-cell IFN- γ ELISpot positive sequences and H2 ^d binding motifs
	NTNSGPDDQIGYYRRATR (76-93)
	GPDDQIGYYRRATRRVRGGD (80-99)
	QIGYYRRATRRVRGGDGK (84-101)
	TRRVRGGDGKMKELSPRW (92-109)
	GKMKELSPRWYFYLY (100-114)
E ^d	GYRRATRR (86-94)
	RRVRGGDGK (93-101)
	VRGGDGKMK (95-103)
	GKMKELSPR (100-108)
	LSPRWYFYLY (105-113)
A ^d	GYRRATRR (86-94)
K ^d	YYRRATRRV (87-95)
	ELSPRWYFYLY (104-112)
L ^d	GPDDQIGYY (80-88)
	LSPRWYFYLY (105-113)
	SPRWYFYLY (106-114)
D ^d	SGPDDQIGY (79-87)
	IGYYRRATR (85-93)
	GYRRATRR (86-94)
	GGDGKMKEL (97-105)
	LSPRWYFYLY (105-113)

61

C)

PRED ^{BALB/c} H2 ^d alleles	N130-149 T-cell IFN- γ ELISpot positive sequences and H2 ^d binding motifs
	GIVWVATEGALNTPKDHIGT (130-149)
K ^d	ALNTPKDHI (139-147)
L ^d	VWVATEGAL (132-140)
A ^d	EGALNTPKD (137-145)
E ^d	WVATEGALN (133-141)
	TEGALNTPK (136-144)

D)

PRED ^{BALB/c} H2 ^d alleles	N241-258 T-cell IFN- γ ELISpot positive sequences and H2 ^d binding motifs
	QQQGQTVTKKSAAEASKK (241-258)
A ^d	QQQGQTVTKK (242-250)
	GQTVTKKSA (244-252)
	VTKKSAAEA (247-255)
	TKKSAAEAS (248-256)

E)

PRED ^{BALB/c} H2 ^d alleles	N352-370 T-cell IFN- γ ELISpot positive sequences and H2 ^d binding motifs
	ILLNKHIDAYKTFPP (352-366)
	LLNKHIDAYKTFPTEPK (353-370)
K ^d	LLNKHIDAY (353-361)
A ^d	NKHIDAYKT (355-363)
E ^d	LLNKHIDAY (353-361)

3.3.4 Performance comparison with other online systems

In Table 3.4, the top 5% of H2^d binders in SARS-CoV N protein predicted by RANKPEP were aligned with the IFN- γ ELISpot positive peptide sequences. Comparing to the top 5% of PRED^{BALB/c} predictions (Table 3.3), RANKPEP picked up 13 fewer positive 9-mers (Table 3.5). RANKPEP failed to pick up any 9-mer binder in two IFN- γ ELISpot positive regions, N76-93 and N92-109 (Table 3.4B), whereas PRED^{BALB/c} predictions picked up three and four 9-mers in the two regions, respectively (Table 3.3B). The IFN- γ ELISpot positive region N76-114 covers three predicted 9-mer binding peptides by the

top 5% of RANKPEP predictions comparing to 16 by the top 5% of PRED^{BALB/c} predictions. The two systems perform comparably in predicting the three other IFN- γ ELISpot positive regions.

Table 3.4 The top 5% of RANKPEP predictions vs. T-cell IFN- γ ELISpot peptides of SARS-CoV N. RANKPEP predicted seven out of nine T-cell IFN- γ immunodominant ELISpot positive sequences from SARS-CoV N. (A) The T-cell IFN- γ ELISpot peptide sequences of SARS-CoV N are indicated by the position number, sequence and RANKPEP alleles predicted. (B) The IFN- γ ELISpot peptides spanning the region N76-114 contained three 9-mer peptides that were within the top 5% of predictions of K^d and L^d alleles. (C) The IFN- γ ELISpot peptides spanning the region N130-149 contained six 9-mer peptides that were within the top 5% of predictions of K^d, A^d, and E^d alleles. (D) The IFN- γ ELISpot peptides spanning the region N241-258 contained four 9-mer peptides that were within the top 5% of predictions of K^d and A^d alleles. (E) The IFN- γ ELISpot peptides spanning the region N352-370 contained two 9-mer peptides that were within the top 5% of predictions of K^d and E^d alleles.

A)

Position	SARS-CoV N ELISpot positive peptide sequences	RANKPEP alleles predicted
76-93	NTNSGPDDQIGYYRRATR	
80-99	GPDDQIGYYRRATRRVRGGD	K ^d
84-101	QIGYYRRATRRVRGGDGK	K ^d
92-109	TRRVRRGGDGKMKELSPRW	
100-114	GKMKELSPRWYFYLY	L ^d
130-149	GIVWVATEGALNTPKDHIGT	K ^d , A ^d , E ^d
241-258	QQQGQTVTKKSAAEASKK	K ^d , A ^d
352-366	ILLNKHIDAYKTFPP	E ^d
353-370	LLNKHIDAYKTFPTEPK	K ^d , E ^d

B)

RANKPEP H2 ^d alleles	N76-114 T-cell IFN- γ ELISpot positive sequences and H2 ^d binding motifs
	NTNSGPDDQIGYYRRATR (76-93)
	GPDDQIGYYRRATRRVRGGD (80-99)
	QIGYYRRATRRVRGGDGK (84-101)
	TRRVRRGGDGKMKELSPRW (92-109)
	GKMKELSPRWYFYLY (100-114)
K ^d	GYRRATR (86-94)
	YYRRATRRV (87-95)
L ^d	SPRWYFYLY (106-114)

63

C)

RANKPEP H2 ^d alleles	N130-149 T-cell IFN- γ ELISpot positive sequences and H2 ^d binding motifs
	GIVWVATEGALNTPKDHIGT (130-149)
K ^d	VWVATEGAL (132-140)
	ALNTPKDHI (139-147)
A ^d	EGALNTPKD (137-145)
E ^d	GIVWVATEG (130-138)
	WVATEGALN (133-141)
	TEGALNTPK (136-144)

D)

RANKPEP H2 ^d alleles	N241-258 T-cell IFN- γ ELISpot positive sequences and H2 ^d binding motifs
	QQQGQTVTKKSAAEASKK (241-258)
K ^d	KKSAAEASK (249-257)
A ^d	QGQTVTKKS (243-251)
	QTVTKKSAA (245-253)
	TKKSAAEAS (248-256)

E)

RANKPEP H2 ^d alleles	N352-370 T-cell IFN- γ ELISpot positive sequences and H2 ^d binding motifs
	ILLNKHIDAYKTFPP (352-366)
	LLNKHIDAYKTFPPTEPK (353-370)
K ^d	AYKTFPPTE (360-368)
E ^d	LLNKHIDAY (353-361)

Table 3.5 Number of predicted 9-mers which fall in the IFN- γ ELISpot positive regions according to the top 5% of predictions of RANKPEP and PRED^{BALB/c}.

	PRED ^{BALB/c}	RANKPEP
Number of 9-mers predicted the using top 5% of predictions	28	15

A large experimental dataset of quantitative MHC-peptide binding data was recently made available [178]. Using this dataset, the performance of different bioinformatics approaches, including PRED^{BALB/c}, SYFPEITHI, RANKPEP, and libscore, in predicting MHC class I binding peptides was compared. The performance of these prediction models has been reported in Table 2 and 3 of the paper [178]. For prediction of 9-mer peptides binding to H2-D^d, PRED^{BALB/c} (AROC=0.929) outperformed all other prediction

tools including the three prediction models developed by Peters *et al.* For prediction of 9-mer peptides binding to H2-K^d, the best performer is the ANN model developed by Peters *et al.* (AROC=0.939) and SYFPEITHI is the best external tool (AROC=0.903). For prediction of 9-mer peptides binding to H2-L^d, the best performer is libscore (AROC=0.987). The AROC values of H2-K^d and -L^d models in PRED^{BALB/c} are 0.807 and 0.934, respectively. The prediction accuracy of PRED^{BALB/c} K^d and L^d models may be limited by the small training datasets (K^d and L^d training datasets contain 244 and 164 peptides, respectively).

3.4 Discussion

No predictive method of peptide-MHC binding consistently outperforms the rest and the most appropriate predictive model depends on the amount of data available [26]. As limited number of training peptides are available for building H2^d models (around 200 peptides available for K^d and L^d molecules), the matrix method has been adopted as prediction method. The prediction models of PRED^{BALB/c} were rigorously tested by 10-fold cross validation, by experimentally identified binders from several protein antigens and by in vivo studies in mice with viral proteins. The accuracy of prediction was excellent (AROC $\geq$ 0.90) for peptide binding to E^d, A^d and K^d molecules and was good (AROC $\geq$ 0.85) for D^d and L^d in 10-fold cross validation. For illustration of its utility we applied PRED^{BALB/c} models in prediction of T-cell epitope sequences in SARS-CoV N protein. The predictions show strong correlation to the experimental IFN- γ ELISpot responses. The top 5% of predictions of H2^d alleles covered all of the nine immunodominant T-cell IFN- γ ELISpot positive epitopes of SARS-CoV N immunogens.

These results indicate that combined analysis of both class I and class II alleles could identify the immunodominant T-cell epitope regions. Although the models accurately predict potential targets of the immune response, the system currently does not always correlate the magnitude of the biological response to the predicted binding strength. This confirms the understanding that the peptide-specific T-cell responses are subject to mechanisms in addition to peptide binding to MHC molecules, including peptide sequence, peptide context, antigen processing, T-cell receptor context, and T-cell repertoire.

We also developed another online prediction system, PRED^{NOD}, for identification of peptides that bind the MHC molecules constituting the H-2^{g7} haplotype of non-obese diabetic (NOD) mice [99]. The H-2^{g7} haplotype comprises the MHC class-I molecules K^d and D^b and a single class-II molecule I-A^{g7}. The NOD mouse is a widely used animal model for study of autoimmune diseases, in particular human type 1 diabetes mellitus (T1DM). Identification of the subset of peptides that bind MHC molecules comprising the H-2^{g7} haplotype and thereby representing potential NOD T-cell epitopes is important for research into the pathogenesis and immunotherapy of T1DM. The derivation of a scoring matrix for I-A^{g7} was difficult due to the small number of experimentally validated non-binders and discrepancies or inconsistencies between experimental peptide datasets determined in various studies. This difficulty was resolved by using an evolutionary algorithm and producing a binding matrix that predicts well across disparate datasets [44]. When tested using an independent dataset, the AROC values were 0.88, 0.96 and 0.96 for predictions of peptide binding to I-A^{g7}, K^d, and D^b molecules, respectively.

Prediction systems like PRED^{BALB/c} and PRED^{NOD} mark a new direction in predictive

modeling of MHC-binding peptides and T-cell epitopes. The main advantage is that PRED^{BALB/c} and PRED^{NOD} focus on a complete organism and their predictions represent a complete set of predicted targets of T-cell immune responses. The focus on the complete set of MHC alleles is closer to studies involving laboratory animals. This approach provides a more complete view of the immune responses of an organism. The two systems are useful for the analysis of immunization regimens and the study of autoimmunity or other immunological diseases.

Chapter 4 PRED^{TAP}: a system for prediction of peptide binding to the human transporter associated with antigen processing

The transporter associated with antigen processing (TAP) is a critical component of the MHC class I antigen processing and presentation pathway. TAP transports antigenic peptides into the ER where it loads them into the binding groove of MHC class I molecules. Because most peptides are first transported by TAP before they are presented on MHC class I molecules, TAP binding preferences should impact significantly on T-cell epitope selection.

4.1 Introduction

Peptides that bind MHC class I molecules serve as recognition targets for CTLs. The major function of CTLs is recognition and destruction of infected (e.g. viruses), mutated (e.g. cancer), or foreign (e.g. transplants) cells. CTLs recognize short antigenic peptides (T-cell epitopes) presented by MHC class I molecules that mainly originate from degradation of cytosolic proteins. Intracellular antigen processing pathways determine the range of peptides which are available for binding to MHC class I molecules and are thereby important factors of CTL responses [179].

MHC class I antigen processing pathway steps include proteasomal cleavage of proteins into shorter peptides, translocation of peptides into the ER by TAP, optional ER trimming by aminopeptidases, insertion of peptides into the binding groove of MHC molecules, and transport of peptide/MHC complexes to the cell surface for presentation to CTLs [5]. TAP is a transmembrane protein responsible for the transport of antigenic

peptides into the ER. TAP demonstrates peptide binding selectivity and the affinity of a particular peptide for TAP influences the probability of its presentation by MHC class I molecules. Peptides that are 8-16 amino acid long and have sufficient binding affinity are efficiently translocated by TAP into the ER, while longer peptides may be transported but with lower efficiency [147]. Human TAP (hTAP) is a heterodimer that has two subunits hTAP1 and hTAP2. TAP belongs to the ATP-binding cassette transporters and each subunit protein has one transmembrane domain and one ATP-binding domain. The genes for human TAP1 and TAP2 are located in the MHC II locus of chromosome 6 and comprise 10 kb each [180]. A more detailed description of function, structure, expression of TAP can be found in [181].

The efficiency of TAP-mediated translocation of a peptide is proportional to its TAP-binding affinity [182, 183]. Mutations, such as premature stop codons, or deletions of either hTAP1 or hTAP2 impair peptide transport into the ER and result in a significant reduction of surface expression of peptide/MHC complexes [184]. TAP deficient cells have low cell-surface HLA class I expression, shown to range from 10% (HLA-A2) to 3%, (HLA-B27 and -A3) [185]. The majority of the peptides presented by HLA class I molecules on cell surface are thus dependent on TAP.

Computational models are routinely used for pre-screening of potential T-cell epitopes and minimization of the number of necessary experiments. Most developments have focused on modeling and prediction of peptide binding to MHC molecules [186]. Amongst computational models of peptide binding to hTAP that have been developed are binding motifs [183], quantitative matrices [187-189], ANN [136, 187], and SVM [128]. As previously shown in Table 2.2, two publicly available online systems, TAPPred [128]

and SVMTAP [6], are based on SVM for prediction of TAP binding peptides. Combined computational methods that integrate multiple critical steps – proteasome cleavage, TAP transport, and MHC class I binding - have been proposed as a supporting methodology for prediction of high probability targets for therapeutic peptides and vaccines [8]. Several combined computational models of antigen processing and presentation have been reported [6, 103, 133, 134, 146]. Testing results indicate that these predictions produce a lower incidence of false positives and reduce the number of experiments required for identification of T-cell epitopes.

Although some publicly available systems for prediction of TAP binding peptides were developed, their performance was not evaluated and compared thoroughly. We developed PRED^{TAP} , a computational system that predicts peptides binding to hTAP. It uses ANN and HMM as predictive engines. Its prediction performance has been compared with those of TAPPred and SVMTAP. PRED^{TAP} is accessible at <http://antigen.i2r.a-star.edu.sg/predTAP/>. The user instructions for PRED^{TAP} are in Appendix B.

4.2 Materials and Methods

4.2.1 Training dataset

There are 493 9-mer peptides in the training dataset (Table 4.1) [136, 187]. In the data file, there are two columns: the first column contains 9-mer peptides and the second column contains the corresponding binding scores. The binding scores range from zero to ten. Scores 7-10 denote high peptide/TAP binding affinity, 5-6 moderate binding affinity, 3-4 low binding affinity and scores 0-2 denote non-binding. The training dataset is available

for download at <http://antigen.i2r.a-star.edu.sg/predTAP/data>.

Table 4.1 Number of peptides in the training dataset.

Binding Affinity	Number of peptides
0	26
1	52
2	48
3	48
4	53
5	55
6	40
7	87
8	61
9	16
10	7
Sum	493

4.2.2 Prediction engines

Provided that a sufficient number of peptide data are available, ANN is the most useful method to provide predictions with high specificity although at the cost of slightly lower sensitivity [26]. Having relatively large datasets of around 500 9-mer peptides (Table 4.1), we implemented the classification engine using an ANN. The inputs to the ANN were the binary strings representing 9-mer peptides. Three amino acid encoding methods, the conventional sparse, 5-bit string, and Rep 9 encoding methods, have been used. Details of the encoding methods are available in section 2.3.3. The outputs were binding scores ranging from zero to ten. The higher the score, the higher the possibility of the peptide being a TAP binder.

Three-layer BP ANN models (in-house software) were used for the development of the PRED^{TAP} server. The learning method was BP with a sigmoid activation function as shown in equation (2.15). The ANN architectures have been tried were 81-2-1, 81-4-1,

and 81-8-1 when using Rep 9 encoding method, 45-2-1, 45-4-1, and 45-8-1 when using 5-bit string encoding method, 180-1-1 and 180-2-1 when using conventional sparse encoding method. The maximum number of ANN training epochs was set to 300. The value of momentum was 0.5 and of learning rate was 0.2. The training stopping criterion was that either the number of training epochs reached the maximum number or the output error E as shown in equation (2.10) was less than 0.01. The training was repeated for four times, and four sets of weights were obtained. The final predictions were the average of the four predictions calculated using the four sets of weights. Details on the ANN and BP algorithm can be found in section 2.4.2.

A HMM model was also built for prediction of TAP binding peptides. For the HMM model, only the 367 binders in the training dataset were used for training. Details on HMM algorithm can be found in section 2.4.3.

The prediction scores produced by ANN and HMM models were normalized using equation (3.1) described in section 3.2.3.

4.3 Results

An assessment of predictive accuracy was carried out for three subsets of peptide binders:

1) all binders including low, moderate, and high binders were considered as positive samples, and all non-binders as negative samples (referred to as the LMH set); 2) moderate and high binders were considered as positive samples, all non-binder and low binders as negative samples (referred to as the MH set), and 3) only high binders were considered as positive samples, with all other peptides as negative samples (referred to as the H set).

The performance of the PRED^{TAP} models was evaluated by 10-fold cross validations and by independent datasets. The prediction performance of PRED^{TAP} was compared with those of publicly available predictive systems, TAPPred and SVM_{TAP}. The details on cross validation are available in section 2.5.1. The performance measures used were introduced in section 2.5.3.

4.3.1 Ten-fold cross validation

The AROC values of ANN models when using Rep 9, 5-bit string, and sparse encoding methods in 10-fold cross validation are shown in Table 4.2, 4.3 and 4.4 respectively. ANN models using conventional sparse encoding methods with two hidden neurons (180-2-1) performed better than those using Rep 9 and 5-bit string encoding methods with eight hidden neurons (81-8-1 and 45-8-1). Thus conventional sparse encoding method was used to encode amino acids for building ANN models in the following studies.

As shown in Table 4.4, for MH set and H set, ANN models showed excellent performance with AROC values above 0.9. For LMH set, the AROC values of ANN models are above 0.85. ANN with structure 180-2-1 showed slightly better performance than that of ANN with structure 180-1-1. Thus ANN with structure 180-2-1 was adopted in our system. The AROC values of HMM models in 10-fold cross validation are shown in Table 4.5. All models showed very good predictive performance. The performance of HMM model is also good with AROC values above 0.85. The specificity vs. sensitivity plot of the ANN prediction model for prediction can be viewed in Figure 4.1. The specificity vs. sensitivity plot of the HMM prediction model can be viewed in Figure 4.2.

Table 4.2 Performance assessment of ANN models when using Rep 9 encoding method in 10-fold cross validation.

ANN 81-2-1	H	MH	LMH
1st run	0.75	0.77	0.72
2nd run	0.82	0.84	0.79
3rd run	0.89	0.90	0.83
ANN 81-4-1	H	MH	LMH
1st run	0.88	0.89	0.83
2nd run	0.90	0.91	0.84
3rd run	0.90	0.90	0.83
ANN 81-8-1	H	MH	LMH
1st run	0.92	0.93	0.84
2nd run	0.92	0.92	0.84
3rd run	0.91	0.91	0.85

Table 4.3 Performance assessment of ANN models when using 5-bit string encoding method in 10-fold cross validation.

ANN 45-2-1	H	MH	LMH
1st run	0.69	0.68	0.64
2nd run	0.77	0.77	0.74
3rd run	0.77	0.76	0.73
ANN 45-4-1	H	MH	LMH
1st run	0.86	0.83	0.80
2nd run	0.85	0.83	0.79
3rd run	0.85	0.83	0.76
ANN 45-8-1	H	MH	LMH
1st run	0.88	0.86	0.80
2nd run	0.85	0.83	0.80
3rd run	0.85	0.82	0.79

Table 4.4 Performance assessment of ANN models when using conventional sparse encoding method in 10-fold cross validation.

ANN 180-2-1	H	MH	LMH
1st run	0.95	0.95	0.89
2nd run	0.95	0.94	0.88
3rd run	0.95	0.94	0.88
ANN 180-1-1	H	MH	LMH
1st run	0.93	0.94	0.87
2nd run	0.92	0.92	0.86
3rd run	0.93	0.94	0.88

Table 4.5 Performance assessment of HMM models in 10-fold cross validation.

HMM	H	MH	LMH
1st run	0.9	0.9	0.87
2nd run	0.89	0.9	0.87
3rd run	0.89	0.9	0.86

Table 4.6 Sensitivities and specificities of ANN and HMM models at the selection threshold 6.0.

ANN 180-2-1	SE	SP
LMH	0.50	1.00
MH	0.67	0.97
H	0.88	0.89
HMM	SE	SP
LMH	0.66	.86
MH	0.81	0.81
H	0.91	0.68

A)

75

B)

C)

Figure 4.1 Sensitivities vs. specificities of the ANN models in PRED^{TAP} obtained from 10-fold cross validation. A) H set; B) MH set; C) LMH set.

76

A)

B)

77

C)

Figure 4.2 Sensitivities vs. specificities of the HMM models in PRED^{TAP} obtained from 10-fold cross validation. A) H set; B) MH set; C) LMH set.

A)

78

B)

C)

Figure 4.3 Sensitivity and specificity of ANN model against thresholds in 10-fold cross validation. The ANN model for prediction of A) H set, B) MH set, and C) LMH set.

79

A)

B)

80

C)

Figure 4.4 Sensitivity and specificity of HMM model against thresholds in 10-fold cross validation. The HMM model for prediction of A) H set, B) MH set, and C) LMH set.

Sensitivities and specificities of the ANN and HMM models at various thresholds (based on normalized scores) in the 10-fold cross validation experiments are shown in Figure 4.3 and Figure 4.4. We selected the normalized score of 6.0 as a reasonable selection threshold, with peptides with scores ≥ 6.0 being predicted TAP binders. In Table 4.6, the sensitivities and specificities of ANN and HMM models at the selection threshold 6.0 are shown. ANN model managed to correctly predict 88% of high binders at the cost of 11% of non-binders, moderate, and low-affinity binders being mistaken as binders; 67% of moderate and high binders with 3% of non-binders and low-affinity binders being mistaken as moderate and high binders; and 50% of all binders (low, moderate, and high) with practically no false positives. The specificities of ANN model for all three sets (LMH, MH, and H sets) are high (1.00, 0.97, and 0.89), which indicates that 6.0 is a

stringent selection threshold and the false positive rate is very low at this threshold. At threshold 6.0, the HMM model managed to correctly predict 91% of high binders with 32% of non-binders, moderate and low-affinity binders being mistaken as high binders; 81% of moderate and high binders with 19% of non-binders and low-affinity binders being mistaken as moderate or high binders; and 66% of all binders (low, moderate, and high) with 14% of non-binders being mistaken as binders. The specificity of the HMM model for LMH set was 0.86, higher than that of the MH set which was 0.81. The specificity of the HMM model for the MH set is much higher than that of the H set, which was 0.68. It implies that the HMM model was able to select binders (low, moderate, and high binders) with low false positive rate, but it failed to categorize them into subgroups - low, moderate, or high binders.

4.3.2 Model validation using independent datasets

To evaluate the predictive power of the methods, the dataset described in Table 4.1 was partitioned into a training set containing two thirds of the data points randomly selected and a testing set containing the remaining one third of data points. The tests were conducted three times for each of the ANN and HMM methods. The AROC values of the ANN and HMM models are shown in Table 4.7. Despite smaller training datasets being used, the ANN models continued to show excellent performance with AROC values above 0.9 for the H and MH sets and good performance with AROC values above 0.85 for the LMH set. The performance of the HMM model is also good with AROC values above 0.85. The performance of HMM dropped slightly with AROC values above 0.85 for the H and MH sets and above 0.80 for the LMH set.

Table 4.7 Performance assessment of ANN/HMM models when the dataset was partitioned into two parts with the training dataset containing two thirds of the data points randomly selected and the testing set containing the remaining one third of data points.

ANN 180-2-1	H	MH	LMH
1st run	0.91	0.92	0.85
2nd run	0.96	0.95	0.90
3rd run	0.94	0.91	0.87
HMM	H	MH	LMH
1st run	0.88	0.88	0.86
2nd run	0.86	0.88	0.83
3rd run	0.91	0.9	0.82

4.3.2 Performance comparison to other predictive systems

The prediction performance of $\text{PRED}^{\text{TAP}}(\text{ANN} \ \& \ \text{HMM})$ was compared with those of publicly available predictive systems, TAPPred (SVM & cascade SVM) [128] and SVMTAP [6]. Two proteins, human papillomavirus (HPV) type 16 E6 (NCBI Accession: P03126) with experimentally identified HLA-A3 binders and human cancer antigen KM-HN-1 (NCBI Accession: NP_689988.1) with three HLA-A24 restricted T-cell epitopes [190], were used and the predicted TAP binders were compared with the HLA binders.

Since PRED^{TAP} , TAPPred, and SVMTAP were built using the same set of training data [136, 187], independent datasets must be used to test and compare their prediction performance. The positions of the top 5% of the predicted TAP binders and HLA-A*0301 binders in HPV type 16 E6 are shown in Table 4.8. SVMTAP, TAPPred (SVM), and $\text{PRED}^{\text{TAP}}(\text{ANN} \ \& \ \text{HMM})$ predicted overlapping sets of peptides, while TAPPred (cascade SVM) predicted very different peptides as compared to other four models. SVMTAP and $\text{PRED}^{\text{TAP}}(\text{ANN})$ produced four identical predictions out of eight predictions with all four of them within the 16-mers containing HLA-A*0301 binders.

SVMTAP and PRED^{TAP} (HMM) produced three identical predictions out of eight predictions with all three of them within the 16-mers containing HLA-A*0301 binders. More than half of the predicted TAP binders were within the 16-mers containing HLA-A*0301 binders in HPV type 16 E6 protein. It was in agreement with the previous study [136] suggesting that HLA-A3 binding peptides have high affinity to TAP.

Table 4.8 The start positions of the top 5% of the predicted TAP binders in HPV type 16 E6 by SVMTAP, TAPPred, and PRED^{TAP}. The positions highlighted in pink were selected by four prediction models. The positions highlighted in blue were those selected by three prediction models. The experimentally identified HLA-A*0301 binders are ¹7–15, ²33–41, ³42–50, ⁴59–67, ⁵75–83, ⁶89–97, ⁷93–101, ⁸125–133, and ⁹143–151. The predictions in the table marked by ^{1–9} are within the 16-mers containing respective HLA-A*0301 binders.

SVMTAP	TAPPred (SVM)	TAPPred (Cascade SVM)	PRED ^{TAP} (ANN)	PRED ^{TAP} (HMM)
75 ⁵	53 ⁴	5 ¹	75 ⁵	75 ⁵
131 ⁸	68	60 ⁴	53 ⁴	46 ³
53 ⁴	80 ⁵	49 ³	46 ³	61 ⁴
150 ⁹	81 ⁵	132 ⁸	68	47 ³
130 ⁸	75 ⁵	93 ⁷	83	7 ¹
46 ³	131 ⁸	116	59 ⁴	53 ⁴
134	134	67	42 ³	49 ³
146 ⁹	51	40 ^{2,3}	130 ⁸	83 ⁶

Three naturally processed peptides from tumor antigen KM-HM-1, namely 196–204, 499–508, and 770–778, are naturally processed by HLA-A*2402 [190]. HLA-A24 binding peptides have been reported as TAP efficient [136, 185]. As KM-HN-1 protein is 833 amino acid long, only the top 2% of the predicted TAP binders (17 entries) are shown in Table 4.9. Peptide 195-203, which is only one amino acid away from the HLA-A*2402 binder 196-204, was picked up by SVMTAP, TAPPred (SVM) and PRED^{TAP} (ANN & HMM). Peptide 506-514, which is within a 16-mer containing the HLA-A*2402 binder 499-508, was picked up by SVMTAP, TAPPred (SVM) and PRED^{TAP} (ANN & HMM).

It was shown that some peptides are efficiently transported by TAP in their optimal size for MHC class I binding, while some peptides are transported as larger peptides that need further trimming in the ER for MHC class I binding [191]. It is possible that peptide 195-204 and 506-514 in KM-HM-1 are transported to ER as longer peptides and further trimmed to 196-204 and 499-508 for loading up to HLA-A24 molecules.

Table 4.9 The start positions of the top 2% of the predicted TAP binders in KM-HN-1 by SVM-TAP, TAPPred, and PRED^{TAP}. The positions highlighted in pink were selected by four prediction models. The positions highlighted in blue were those selected by three prediction models. The experimentally identified HLA-A*2402 binders are ¹196–204, ²499–508, and ³770–778. The predictions in the table marked by ^{1–3} are within the 16-mers containing respective HLA-A24 binders.

SVMTAP	TAPPred (SVM)	TAPPred (Cascade SVM)	PRED ^{TAP} (ANN)	PRED ^{TAP} (HMM)
Position	Position	Position	Position	Position
660	372	674	195 ¹	682
372	195 ¹	314	654	506 ²
426	426	639	372	372
195 ¹	794	249	422	683
794	330	530	565	507
199	317	525	317	195 ¹
654	660	325	310	492
317	331	206	378	310
371	652	12	468	660
110	199	479	763 ³	16
760	198	537	337	468
198	371	112	426	317
789	654	93	737	395
705	506 ²	626	246	573
457	789	470	660	223
507	457	141	756	193
48	730	483	110	730
573	304	71	506 ²	313
376	565	99	201	647
652	395	668	507	510
395	760	781	365	652
780	318	57	653	15
455	764 ³	579	502 ²	814
506 ²	63	124	492	676
324	507	590	456	782

4.4 Discussion

In training of ANN, three amino acid encoding methods, the conventional sparse, 5-bit string, and Rep 9 encoding methods, have been tried. Previously it was reported that little difference on prediction performance was observed from using Rep 6 and Rep 9 amino acid encoding methods and the sparse encoding method [141]. In our experiments, ANN models using the sparse encoding methods with two hidden neurons (180-2-1) performed better than those using Rep 9 and 5-bit string encoding methods with eight hidden neurons (81-8-1 and 45-8-1). Thus we adopted conventional sparse encoding method for encoding of amino acids.

The prediction models of PRED^{TAP} were rigorously tested by 10-fold cross validation and by independent datasets. The accuracy of prediction was good (AROC ≥ 0.85) for ANN and HMM models in 10-fold cross validation. In an earlier study, four computational servers for prediction of H-2K^d binding peptides were compared with each other [20]. A 121-amino acid long sequence of the nuclear export protein NS2 from influenza A virus (GenPept Accession: NP_859033) was searched for 9-mer candidate binders to a mouse MHC molecule H-2K^d using four online systems. Only three peptides were predicted within the top ten candidates as binders by all four methods. The performance comparison of PRED^{TAP} with SVMTAP and TAPPred (SVM) shows that consensus peptides can be selected by combining predictions. The examples suggested that individual predictions need to be taken with care and predictions may be improved by a consensus of multiple methods. A similar situation may be applicable to TAP predictions. Hence the combination of ANN and HMM predictions in PRED^{TAP} should result in higher specificity (fewer false positives) at the cost of slightly lower sensitivity.

Combined computational methods that integrate multiple critical steps of the MHC class I antigen processing pathway have been proposed as a supporting methodology for prediction of T-cell epitopes and several applications modeling the pathway have been reported. However, these combined predictions need to be taken with a dose of caution. Alternative pathways for both proteolytic degradation [192] and TAP transport [193] have been reported. In some cases TAP-deficient individuals have normal immune responses [194], suggesting that TAP-independent immune responses are sufficient to provide effective protection from some intracellular pathogens. Nevertheless, the proteasome-TAP-MHC class I pathway is responsible for 90-97% of expression of peptide/MHC Class I complexes and therefore is critical for the identification of target epitopes for immunotherapies and vaccines.

The combinatorial properties of molecular mechanisms involved in antigen processing and the adaptive learning nature of the immune responses limit our ability to fully predict immune responses. Combining experimental and computational techniques improves our ability to decipher complex interactions of the immune system. Computer models are used to complement laboratory experiments and thereby speed up knowledge discovery in immunology. In particular, the number of large-scale laboratory experiments for T-cell epitope mapping can be minimized by the judicious use of experiments aimed at developing and validating computer models. These models can then be used to perform large-scale computer simulations rapidly and inexpensively. The hypotheses generated from these experiments can then be retested in the laboratory to confirm their applicability to real-life immunology. PRED^{TAP} provides for the prediction of peptide binding by TAP and can be used as a comparison method against other TAP prediction

87

servers.

Chapter 5 MULTIPRED1: a computational system for prediction of promiscuous HLA binding peptides

The process of establishing a prediction system of peptide binding to MHC molecules involves three major steps as shown in Figure 5.1. The first step is collection of a list of peptides with known binding affinities to a particular MHC molecule and conversion of these data to a format suitable for input to a computer program for generating training and test data for prediction system. The second step involves defining algorithms and their parameters, training the model using the training data and testing the models built using a completely independent test dataset. If the test results are not satisfactory, go back to model building again with redefined parameters and repeat the steps until the model is of acceptable accuracy. After a prediction model is built and tested, it is ready for prediction of MHC binding peptides from input proteins.

Each of the steps of building models for prediction of MHC binders is described in detail in this chapter. Datasets acquired are introduced, followed by a novel data representation method proposed to enable predictions of peptides binding to multiple related HLA alleles belonging to one supertype. The parameters used in prediction engines, ANN, HMM and SVM, are introduced. The performance of models built using the three algorithms are compared and reported. We also report an online prediction system MULTIPRED1 that is able to predict peptide binding to multiple HLA molecules belonging to one supertype using a single model.

Figure 5.1 Steps involved in building a prediction system.

5.1 Introduction

An ideal epitope-based vaccine needs to contain minimal number of epitopes that cover a vast majority of human population [135, 195]. Promiscuous peptides that bind more than one HLA allele are prime targets for vaccine and immunotherapy development because they are relevant to higher proportions of the human population than peptides that bind single alleles. Several methods and prediction systems have been developed for the prediction of promiscuous peptides.

A module concept for HLA class I peptide binding prediction was proposed [130], which implements a data representation method capturing peptide/HLA interaction information. The HLA binding groove that interacts with a specific position in the bound peptide is known as a pocket. The contact positions in HLA binding grooves were provided by [196]. A module is the sequence of contact positions in a specific HLA allele for an amino acid in the bound peptide. The modular concept increased the number of predictable alleles from 15 to 75 of HLA-A and 12 to 36 of HLA-B molecules. Under the modular concept, binding data of certain HLA alleles can make prediction possible for numerous additional alleles. PeptideCheck is the online prediction system based on the module concept and matrix method. The module concept is very similar to the data representation method proposed in [157, 197], which captures the interaction between a peptide and a HLA binding groove based on the binding environment defined by [196]. The data representation method makes it possible to predict promiscuous peptides binding to multiple MHC alleles belonging to a supertype using one single prediction model.

Several other prediction systems, such as PEPVAC [135] and NetCTL [134] have been developed to identify promiscuous HLA binding peptides. PEPVAC identifies promiscuous peptide binders to five HLA class I supertypes (A2, A3, B7, A24, and B15) using PSSM in combination with a dynamic algorithm. The estimated frequency of phenotypic population of these supertypes is 95%, regardless of ethnicity. NetCTL employs an integrative approach to CTL epitope prediction which combines predictions of MHC class I binding, TAP transport efficiency, and proteasomal cleavage. It predicts peptides binding to 12 HLA class I supertypes, A1, A2, A3, A24, A26, B7, B8, B27, B39, BB44, B58, and B62.

We developed MULTIPRED1, a web-based system for prediction of promiscuous peptides that bind multiple HLA alleles. A novel data representation method enables MULTIPRED1 to predict peptides that bind to multiple HLA alleles belonging to one HLA supertype by a single prediction model per supertype. It can also predict peptide binding to HLA alleles for which no experimental binding data are currently available. The current implementation can predict peptides that bind HLA proteins belonging to supertypes A2, A3 and B7 (HLA class I) and DR (HLA class II). A2, A3 and B7 supertypes cover approximately 88% of human population, irrespective of ethnicity [135], while class II DR supertype is present in 100% of human population [198]. The predictive engines implemented in MULTIPRED are ANN, HMM, and SVM [157, 197]. MULTIPRED1 is available at <http://antigen.i2r.a-star.edu.sg/multipred1/>. An earlier version of MULTIPRED1, MULTIPRED [199], is available at <http://antigen.i2r.a-star.edu.sg/multipred/>.

5.2 Materials and Methods

5.2.1 Training datasets

The predominant length of peptides that bind HLA class I molecules (HLA-A, -B, and -C) is nine amino acids [174]. Thus, only 9-mer peptides were used in building our prediction models. HLA-A2 supertype alleles are estimated to be present in more than 40% of the world population [200] with A*0201 being the most common HLA-A2 supertype allele. HLA-A3 supertype alleles are estimated to be present in more than 35% of the world population [200]. HLA-B7 supertype alleles are estimated to be present in more than 40%

of the world population [200]. B*0702, B*3501, B*5101, B*5301 and B*5401 are the most common HLA-B7 supertype alleles.

Table 5.1 Number of peptides binding or non-binding to 15 HLA-A2 alleles in Dataset1.

HLA-A2 alleles	Number of peptides		
	Binders	Non-binders	Sum
*0201	451	1988	2439
*0202	45	25	70
*0203	46	7	53
*0204	23	224	247
*0205	16	40	56
*0206	43	37	80
*0207	4	11	15
*0208	0	4	4
*0209	5	1	6
*0210	3	0	3
*0211	4	0	4
*0214	8	1	9
*0217	2	4	6
*6802	23	31	54
*6901	2	2	4
All	675	2375	3050

Table 5.2 Number of peptides binding or non-binding to eight HLA-A3 alleles in Dataset2.

HLA-A3 alleles	Number of peptides		
	Binders	Non-binders	Sum
*A0301	107	89	196
*A0302	146	259	405
*A1101	142	223	365
*A1102	142	211	353
*A3101	44	54	98
*A3301	35	62	97

*A3303	5	0	
*A6801	59	638	697
All	680	1536	2216

Table 5.3 Number of peptides binding or non-binding to 13 HLA-B7 alleles in Dataset3.

HLA-B7 alleles	Number of peptides		
	Binders	Non-binders	Sum
*0702	301	1273	1574
*1508	6	0	6
*3501	373	779	1152
*3502	13	1	14
*3503	18	2	20
*5101	166	176	342
*5102	27	5	32
*5103	28	5	33
*5301	139	211	350
*5401	97	188	285
*5501	82	204	286
*5502	5	0	5
*5601	3	0	3
All	1258	2844	4102

The HLA-A2 and A3 training data are from three sources, MHCPEP database [21], a set of 9-mer non-binders (V. Brusic, unpublished data) and additional published papers. In total 3050 9-mer peptide data (referred to as Dataset1) comprising 675 binders and 2375 non-binders related to 15 alleles of the HLA-A2 supertype (Table 5.1) and 2216 9-mer peptide data (referred to as Dataset2) comprising 680 binders and 1536 non-binders related to eight alleles of the HLA-A3 supertype (Table 5.2) were collected. Also 4102 9-mer peptides (referred to as Dataset3) related to 13 alleles of HLA-B7 supertype (Table 5.3) were collected from databases, e.g. MHCPEP [21], SYFPEITHI [22], EPIMHC [23]

and AntiJen [24], and literature [59, 107, 178, 200-209]. The training datasets are available for download at <http://antigen.i2r.a-star.edu.sg/multipred1/data>.

These data have been collected from multiple sources and often the binding affinities between peptides and MHC molecules were measured differently. The differences may come from various aspects, such as different binding indicator (bound radioactive ligand or quantitative enzyme-linked immunosorbent assay), what is detected (competitive binding or refolding), the way the MHC molecules are prepared (isolated from homozygous cell lines or recombinant MHC), and the purity of peptides [178]. Refer to [210] for a review of the various methods for in vitro measurement of MHC binding peptides. The quality of measurements of peptide binding varies between different experimental methods and also with the MHC molecules studied [211]. Thus inconsistency, biases, and noise have been included in the datasets inevitably when combining data from multiple sources.

5.2.2 Data Preprocessing Method

A novel representation of peptide/HLA interaction, schematically presented in Figure 5.2, was developed. A peptide/HLA interaction is represented by a virtual peptide composed of peptide residues and HLA residues that come in contact with the peptide. The data representation used in building ANN and SVM models is slightly different from that used in building HMM models. For ANN and SVM model building, a virtual peptide is comprised of peptides residues followed by all the contact residues in the receptor (Figure 5.2C). For HMM model building, in a virtual peptide, each residue of the peptide is followed by its contact residues in the receptor (Figure 5.2D). For ANN and SVM

methods, the order of the amino acids in an input peptide does not affect the models built, the shorter peptide representation as shown in Figure 5.2C was thus adopted. For HMM method, the order of the amino acid affects the models built, the peptide representation as shown in Figure 5.2D was used.

Figure 5.2 Building the representation of peptide/receptor interaction. The contact residues are those that are involved in interaction. A) Identification of contact residues in the receptor that are involved in peptide/receptor interaction. B) Linear representation of peptide and receptor. C) Data presentation used in ANN and SVM model building: Removal of non-contact residues and combining peptide and contact residues in receptor to form a virtual peptide representing peptide/receptor interaction. D) Data presentation used in HMM model building: Removal of non-contact residues and each amino acid in the peptide is followed by its contact residues in receptor to form a virtual peptide representing peptide/receptor interaction.

The HLA protein sequences were extracted from the IMGT/HLA Sequence Database, release 2.4.0 (<http://www.anthonynolan.org.uk/HIG/>). To simplify the data representation and eliminate redundant information, we considered only those contact residues that vary across the HLA-A2/A3/B7 alleles appearing in Dataset1/Dataset2/Dataset3, and discarded the residues that are conserved across those HLA-A2/A3/B7 alleles. The contact residues conserved across the alleles will not provide any useful information to feature vectors, as they are same in all feature vectors. There are in total 48 receptor contact residues (Table 5.4) of which 31 are non-conserved across the HLA-A alleles (Table 5.5) and 28 are non-conserved across the HLA-B alleles (Table 5.6) as reported in [196].

Table 5.4 Peptide positional environment for HLA class I molecules. P denotes peptide residues and R denotes HLA contact residues.

Position	Receptor contact residues (R)
P1	R5, R7, R33, R59, R62, R63, R66, R99, R159, R163, R167, R171
P2	R7, R9, R24, R25, R26, R34, R35, R36, R45, R62, R63, R66, R67, R70, R99, R159, R163, R167
P3	R7, R9, R62, R66, R70, R97, R99, R114, R152, R155, R156, R159, R163
P4	R62, R65, R66, R69, R70, R155, R156, R159
P5	R69, R70, R73, R74, R97, R114, R116, R152, R155, R156, R159
P6	R7, R9, R22, R24, R66, R69, R70, R73, R74, R97, R99, R114, R116, R133, R147, R152, R155, R156
P7	R73, R77, R97, R114, R116, R133, R146, R147, R150, R152, R155, R156
P8	R73, R76, R77, R80, R97, R143, R146, R147
P9	R70, R73, R74, R76, R77, R80, R81, R84, R95, R96, R97, R114, R116, R123, R124, R142, R143, R146, R147
All	R5, R7, R9, R22, R24, R25, R26, R33, R34, R35, R36, R45, R59, R62, R63, R65, R66, R67, R69, R70, R73, R74, R76, R77, R80, R81, R84, R95, R96, R97, R99, R114, R116, R123, R124, R133, R142, R143, R146, R147, R150, R152, R155, R156, R159, R163, R167, R171

The tables listing the non-conserved residues for each allele are reported in Table 5.7 to Table 5.9. Out of the 31 non-conserved residues across all HLA-A alleles, 12 are non-conserved across the 15 HLA-A2 alleles in Dataset1 and 12 are non-conserved across the eight HLA-A3 alleles in Dataset2, as shown in Table 5.7 and Table 5.8. Out of the 28 non-conserved residues across all HLA-B alleles, 18 are non-conserved across the 13 HLA-B7 alleles in Dataset3, as shown in Table 5.9.

Each virtual long peptide constructed, representing the interaction between a 9-mer peptide and a HLA molecule, is specific to the HLA molecule. This is done to reflect the variation(s) observed in the contact position(s) among the HLA molecules. In building ANN models for prediction of HLA binders, the representation was based on all HLA-A variants. Out of the 29 non-conserved residues, there are 12 of them are non-conserved across the 15 HLA-A2 alleles as shown in Table 3. By combining the 9-mer peptide and the 12 non-conserved contact residues, a virtual long peptide with 21 amino acids capturing the interaction information between a peptide and a HLA-A2 allele was defined as

P1-P2-P3-P4-P5-P6-P7-P8-P9-R62-R63-R66-R70-R73-R74-R95-R97-R99-R152-R156. P denotes peptide residues and R denotes HLA-A contact residues. For example, the virtual peptides representing a 9-mer peptide, ALMDKSLHV, that binds both HLA*A0201 and *0202 are ALMDKSLHVYFAFRFGERKVAH^{**THVD**}TLVRYHYTAV***LT***WY and ALMDKSLHVYFAFRFGERKVAH^{**THVD**}TL***LL***RYHYTAV***WT***WY, respectively. The two contact positions, R95 and R156 (shown in bold and italic), possess different amino acid residues in HLA*A0201 and *0202, and hence we have generated a virtual peptide

for each of these two HLA molecules. The virtual peptide representation of peptide/HLA-A interactions was encoded as a binary string for use in ANN/SVM training and prediction. Conventional sparse encoding method has been used to encode peptide residues. For HLA (receptor) contact residues, the length of binary string depends on the number of different residues observed at variable contact positions in HLA molecules (Table 5.5). For example R7 has length 2 because there are only 2 possible variants observed in that position, namely “Y” and “C”, and were encoded as 10 and 01, respectively. After combining all HLA contact residues as binary strings, the length of representation of peptide/HLA-A interaction was 275 bits.

In building HMM models, we represented a HLA-peptide interaction by combining each amino acid of the peptide with the variable amino acids of its positional environment. The contact residues were used in the order they appear in the linear sequence of HLA molecules. An example of the representation of a HLA-A3/peptide interaction is a virtual peptide:P1-R62-R63-R163-R171-P2-R9-R62-R63-R70-R163-P3-R9-R62-R70-R97-R114-R152-R156-R163-P4-R62-R70-R156-P5-R70-R73-R97-R114-R152-R156-P6-R9-R70-R73-R97-R114-R152-R156-P7-R73-R97-R114-R152-R156-P8-R73-R97-P9-R70-R73-R97-R114-R142. The virtual peptide contains 54 amino acids both from the peptide and the HLA-A3 molecule. The data preprocessing methods have also been reported in [197, 212].

Table 5.5 Non-conserved residues of all HLA-A alleles.

Contact Position	R7	R9	R24	R33	R35	R36	R62	R63	R65	R66	R67	R69	R70	R73	R74	R76
Possible amino acid	Y C	F S Y T	A S	F L	R Q	F L	Q G E R L	E Q N	R G	N K Q H	M V	A G	H Q	T I S A	D H N	A E V
Contact Position	R77	R80	R81	R95	R97	R99	R114	R116	R142	R150	R152	R156	R163	R167	R171	
Possible amino acid	N D S	T I	L A	I V L	I M R	Y C F S H	R H Q E	D Y N H V S F	I T	V A	A V E W R	R L W Q	R W T E L	G W	Y H	

Table 5.6 Non-conserved residues of all HLA-B alleles.

Contact Position	R9	R24	R45	R59	R62	R63	R65	R66	R67	R69	R70	R74	R76	R77
Possible amino acid	Y D H	S T A P	E G M K T	Y H F	R P G E	N E D	Q R	I K T N	Y F C S M	A T R I G	Q N S K	D Y G	E V	S N D G
Contact Position	R80	R81	R95	R97	R99	R114	R116	R143	R147	R152	R156	R163	R167	R171
Possible amino acid	N T I K	L A P	L I W	S R T W N V M	Y F S C	D N H K	Y S D L F H N	T S I	W L	E V Q A	R D L W E	E T L A G M	W S G	Y H

Table 5.7 Non-conserved contact residues across the 15 HLA-A2 alleles appearing in Dataset1. Variations have been highlighted.

Positions	9	62	63	66	70	73	74	95	97	99	152	156
A*0201	F	G	E	K	H	T	H	V	R	Y	V	L
A*0202	F	G	E	K	H	T	H	L	R	Y	V	W
A*0203	F	G	E	K	H	T	H	V	R	Y	E	W
A*0204	F	G	E	K	H	T	H	V	M	Y	V	L
A*0205	Y	G	E	K	H	T	H	L	R	Y	V	W
A*0206	Y	G	E	K	H	T	H	V	R	Y	V	L
A*0207	F	G	E	K	H	T	H	V	R	C	V	L
A*0208	Y	G	E	N	H	T	H	L	R	Y	V	W
A*0209	F	G	E	K	H	T	H	V	R	Y	V	L
A*0210	Y	G	E	K	H	T	H	V	R	F	V	L
A*0211	F	G	E	K	H	I	D	V	R	Y	V	L
A*0214	Y	G	E	K	H	T	H	L	R	Y	V	L
A*0217	F	G	E	K	H	T	H	L	M	F	V	L
A*6802	Y	R	N	N	Q	T	D	I	R	Y	V	W
A*6901	Y	R	N	N	Q	T	D	V	R	Y	V	L

Table 5.8 Non-conserved contact residues across the eight HLA-A3 alleles appearing in Dataset2. Variations have been highlighted.

Positions	9	62	63	70	73	97	114	142	152	156	163	171
A*0301	F	Q	E	Q	T	I	R	I	E	L	T	Y
A*0302	F	Q	E	Q	T	I	R	I	V	Q	T	Y
A*1101	Y	Q	E	Q	T	I	R	I	A	Q	R	Y
A*1102	Y	Q	E	Q	T	I	R	I	A	Q	R	Y
A*3101	T	Q	E	H	I	M	Q	I	V	L	T	Y
A*3301	T	R	N	H	I	M	Q	I	V	L	T	H
A*3303	T	R	N	H	I	M	Q	I	V	L	T	Y
A*6801	Y	R	N	Q	T	M	R	T	V	W	T	Y

Table 5.9 Non-conserved contact residues across the 13 HLA-B7 alleles in Dataset3. Variations have been highlighted.

	24	45	67	69	70	74	77	80	81	95	97	114	116	152	156	163	167	171
B*0702	S	E	Y	A	Q	D	S	N	L	L	S	D	Y	E	R	E	W	Y
B*1508	A	M	F	T	N	Y	S	N	L	L	R	D	S	E	W	L	W	Y
B*3501	A	T	F	T	N	Y	S	N	L	I	R	D	S	V	L	L	W	Y
B*3502	A	T	F	T	N	Y	S	N	L	I	R	N	Y	V	L	L	W	Y
B*3503	A	T	F	T	N	Y	S	N	L	I	R	D	F	V	L	L	W	Y
B*5101	A	T	F	T	N	Y	N	I	A	W	T	N	Y	E	L	L	W	H
B*5102	A	T	F	T	N	Y	N	I	A	W	T	N	Y	E	L	L	W	Y
B*5103	A	T	F	T	N	Y	N	I	A	W	T	N	Y	E	L	L	G	H
B*5301	A	T	F	T	N	Y	N	I	A	I	R	D	S	V	L	L	W	Y
B*5401	A	G	Y	A	Q	D	S	N	L	W	T	N	L	V	L	T	W	Y
B*5501	A	E	Y	A	Q	D	S	N	L	W	T	N	L	E	L	T	W	Y
B*5502	A	E	Y	A	Q	D	S	N	L	W	T	N	L	V	L	T	W	Y
B*5601	A	E	Y	A	Q	D	S	N	L	W	T	N	L	V	L	L	W	Y

5.2.3 Prediction engines

Having relatively large datasets of 3050, 2216 and 4102 9-mer peptides related to HLA-A2, A3, and B7 supertypes, respectively (Table 5.1, Table 5.2, and Table 5.3), we implemented a classification engine using ANNs. The inputs to the ANN were the binary strings representing virtual peptides describing peptide/HLA interactions, the outputs were binding scores ranging from one to nine. Scores eight and nine denoted high peptide/HLA binding affinity, six and seven moderate binding affinity, four and five low binding affinity and scores less than four denote non-binding.

Training parameters of each ANN model were determined by observing 100 trial runs. The learning method was BP with a sigmoid activation function as shown in equation (2.15). For both A2 and A3 supertypes, the three-layer BP network architecture was 276-4-1, the maximum number of the ANN training epochs was set to 300, the value of momentum was 0.5 and of learning rate was 0.001. The training was repeated four times, and four sets of weights were obtained. For the HLA-B7 supertype, the three-layer BP network architecture was 289-8-1, the maximum number of training epochs was set to 500, the value of momentum was 0.5 and of learning rate 0.005. In building the ANN models for each supertype, the training was repeated four times, and four sets of weights were obtained. The final predictions were the averages of the four predictions calculated using the four sets of weights. Details on ANN and BP algorithm can be found in section 2.4.2.

HMM models have also been built for prediction of HLA binding peptides. For the HMM models, only the binders in Dataset1, Dataset2, and Dataset3 are used for training. Details on the HMM algorithm can be found in section 2.4.3.

The training of SVM was carried out using SVMlight [213]. The input for this package follows the format of feature/value, and we used each input pair as an indicator of whether a particular amino acid is present at a particular position. Various kernel functions were examined (linear, polynomial and radial basis), and parameters were optimized (trade-off C , power d for polynomial kernel, and parameter g for radial kernel). The models that had the best performance (in terms of the highest average blind testing AROC value for molecules within a given supertype) were: Gaussian kernel with $g=0.1$ and $C=0.5$ for HLA-A2, Gaussian kernel with $g=0.1$ and $C=2$ for HLA-A3, and 2-degree polynomial or Gaussian kernel with $g=0.05$ and $C=0.5$ for HLA-B7 supertype. Details on the SVM algorithm can be found in section 2.4.4. In addition to performing binary classification using SVM, regression analysis has also been performed using SVMlight for predictions of peptide/HLA binding affinities.

5.3 Results

An assessment of predictive accuracy was carried out for three subsets of peptide binders:

1) all binders including low, moderate, and high binders were considered as positive samples, and all non-binders as negative samples (referred to as the LMH set); 2) moderate, and high binders were considered as positive samples, all non-binder and low binders as negative samples (referred to as the MH set), and 3) only high binders were considered as positive samples, with all other peptides as negative samples (referred to as the H set).

The performance of the prediction models was evaluated by blind tests, by 10-fold cross validations, and by several sets of experimentally identified HLA binding data. The

prediction performance of MULTIPRED1 was compared to those of several publicly available predictive systems too. The details on cross validation are available in section 2.5.1. The performance measures used were introduced in section 2.5.3.

5.3.1 Blind testing of HLA-A2 ANN models

To test the performance of ANN models for the prediction of HLA-A2 promiscuous peptides, we constructed test sets comprising peptide data of five HLA-A2 alleles. The numbers of peptides in training and testing datasets are shown in Table 5.10. The other ten HLA-A2 alleles in Dataset1 were not considered for testing because of the small number of related peptides. However, all data were used for ANN training. To test the predictive accuracy of peptide binding to each of the five HLA-A2 alleles, we used all peptides (binders and non-binders) related to this allele available in the Dataset1 as the testing data and used all peptides related to other alleles as training data. The prediction of peptides that bind to one HLA-A2 allele was, therefore, done without any prior knowledge of peptide binding to this allele. This type of test is referred to as a blind test in this thesis. The successful blind test results imply that predictions can also be performed for HLA-A2 alleles for which experimental data are not available.

Table 5.10 Blind testing on five HLA-A2 alleles: Number of peptides in the training and testing sets for ANN models.

HLA-A2 alleles	Training data		Testing data	
	Binders	Non-binders	Binders	Non-binders
*0201	224	387	440	1999
*0202	619	2361	45	25
*0204	641	2162	23	224
*0205	648	2346	16	40
*0206	621	2349	43	37

The binders represent only 22% of the Dataset1, while non-binders represent the remaining 78% (Table 5.1). This bias is present in each training dataset shown in Table 5.10. The unbalanced proportion of binders and non-binders in training datasets may affect the training of the ANN and the resulting predictive accuracy. To optimize the dataset composition, multiple training datasets were constructed. For each HLA-A2 allele, ten training sets and one test set were constructed. For example, the HLA-A*0201 test set had 451 binders and 1988 non-binders. The first HLA-A*0201 training set was constructed from Dataset1 and had 224 binders and 387 non-binders. The second HLA-A*0201 training set had two copies of each binder, in total 448 binders (2×224) and 387 non-binders. The third HLA-A*0201 training set had three copies of each binder, in total 672 binders (3×224) and 387 non-binders. The HLA-A*0201 training sets four through ten were constructed by adding one more copy of binders up to 10 copies in the tenth training set. We trained ANN models for each of the five HLA-A2 molecules (*0201, *0202, *0204, *0205, and *0206) using the ten datasets to find the composition of training data that results in best predictive performance of the trained system.

A total of 150 simulations were performed for testing: five HLA-A2 alleles, ten training datasets for each allele, with three runs for each dataset. We calculated 450 AROC values (for H, MH, and LMH sets in each simulation). The AROC values ranged from 0.58 to 0.97 with average of 0.85. The maximum standard deviation for repeated runs was 0.06 and average standard deviation was 0.02. The test results were therefore highly reproducible. Forty-three values (9.6% of the total) represented models that had $0.58 \leq \text{AROC} < 0.7$ and therefore considered not useful. All low-scoring AROC values pertained to the models of HLA-A*0202 and included H or MH sets. The average AROC

values for HLA-A*0201, *0202, *0204, *0205, and *0206 were 0.85, 0.70, 0.9, 0.93, and 0.89, respectively. These results indicate good or excellent predictive performance for the majority of HLA-A2 alleles, except for a marginal predictive performance of the HLA-A*0202 model.

Figure 5.3 Performance comparison of HLA-A2 ANN models trained using training datasets of various compositions.

Table 5.11 AROC values for blind testing of HLA-A2 ANN models, with four copies of binders in the training datasets.

	AROC				
	*0201	*0202	*0204	*0205	*0206
LMH	0.87	0.76	0.88	0.93	0.91
MH	0.87	0.71	0.91	0.95	0.90
H	0.87	0.66	0.90	0.92	0.87

Predictive performance of the models was dependent on the binder/non-binder composition of the training sets. The average AROC values of ANN models trained with the ten training datasets are shown in Figure 5.3. The ANN models trained using datasets containing four copies of binders and six copies of binders gave the overall best performance (Figure 5.3). The test results for the selected optimal model (four copies of binders) are shown in Table 5.11. Predictions for HLA-A*0202 were of marginal accuracy for LMH and for MH sets, but not useful for the H set (high affinity binders). All other models showed very good or excellent predictive performance (AROC>0.85). For models trained by datasets containing six copies of the binders, the AROC values were similar to the results shown in Table 5.11.

5.3.2 Blind testing of HLA-A3 ANN models

To test the performance of ANN models for the prediction of HLA-A3 promiscuous peptides, we did blind tests on seven HLA-A3 alleles. The numbers of peptides in the training and testing datasets are shown in Table 5.12. HLA*A3303 was not considered for testing because there are only five peptides related to it in Dataset2. For details on blind tests, refer to section 5.3.1.

The binders represent only 31% of Dataset2, while non-binders represent the remaining 69% (Table 5.2). The unbalanced proportion of binders and non-binders in training datasets may affect the training of ANN and the resulting predictive accuracy. To optimize dataset composition, multiple training datasets were used. Following the same method of constructing training datasets of various compositions in section 5.3.1, ten training sets and one test set were constructed for each HLA-A3 allele. We trained ANN

models for each of the seven HLA-A3 molecules (*0301, *0302, *1101, *1102, *3101, *3301, and *6801) using the ten datasets to find the composition of training data that results in best predictive performance of the trained system.

Table 5.12 Blind testing on seven HLA-A3 alleles: Numbers of peptides in training and testing sets for ANN models.

HLA-A3 alleles	Training data		Testing data	
	Binders	Non-binders	Binders	Non-binders
*0301	573	1447	107	89
*0302	534	1277	146	259
*1101	538	1313	142	223
*1102	538	1325	142	211
*3101	636	1482	44	54
*3301	645	1474	35	62
*6801	621	898	59	638

A total of 210 simulations were performed for testing: seven HLA-A3 alleles, ten training datasets for each allele, with three runs for each dataset. We calculated 630 AROC values (for H, MH, and LMH sets in each simulation). The AROC values ranged from 0.46 to 0.98 with average of 0.79. The maximum standard deviation for repeated runs was 0.119 and average standard deviation was 0.028. The test results were therefore highly reproducible. One hundred and sixty two values (25.7% of the total) represented models that had $0.46 \leq \text{AROC} < 0.7$ and therefore considered not useful. 159 out of the 162 low-scoring AROC values pertained to the models of HLA-A*3101 and *3301. The average AROC values for HLA-A*0301, *0302, *1101, *1102, *3101, *3301, and *6801 were 0.83, 0.80, 0.89, 0.84, 0.60, 0.62, and 0.93, respectively. These results indicate good

or excellent predictive performance for the majority of HLA-A3 alleles, except for a marginal predictive performance with HLA-A*3101 and *3301 models.

The average AROC values of ANN models trained with the ten training datasets are shown in Figure 5.4. The test results for the selected optimal model (four copies of binders) are shown in Table 5.13. Predictions for HLA-A*3101 and *3301 were of marginal accuracy. All other models showed very good or excellent predictive performance (AROC>0.8).

Figure 5.4 Performance comparison of HLA-A3 ANN models trained using training datasets of various compositions.

Table 5.13 AROC values for blind testing of HLA-A3 ANN models, with four copies of binders in the training datasets.

	AROC						
	*0301	*0302	*1101	*1102	*3101	*3301	*6801
LMH	0.89	0.84	0.91	0.86	0.69	0.63	0.96
MH	0.87	0.82	0.88	0.83	0.60	0.61	0.96
H	0.82	0.81	0.88	0.83	0.54	0.64	0.96

5.3.3 Blind testing of HLA-B7 ANN models

To test the performance of ANN models for the prediction of HLA-B7 promiscuous peptides, we did blind tests on six HLA-B7 alleles. The numbers of peptides in training and testing datasets are shown in Table 5.14. The other seven HLA-B7 alleles in Dataset3 were not considered for testing because of the small number of related peptides. For details on blind tests, refer to section 5.3.1.

Table 5.14 Blind testing on six HLA-B7 alleles: Number of peptides in the training and testing sets for ANN models.

HLA-B7 alleles	Training data		Testing data	
	Binders	Non-binders	Binders	Non-binders
*0702	957	1571	301	1273
*3501	885	2065	373	779
*5101	1092	2668	166	176
*5301	1119	2633	139	211
*5401	1161	2656	97	188
*5501	1176	2640	82	204

The binders represent 31% of Dataset3, while non-binders represent the remaining 69% (Table 5.3). The unbalanced proportion of binders and non-binders in training datasets may affect the training of the ANN and the resulting predictive accuracy. To optimize dataset composition, multiple training datasets were used. Following the same method of constructing training datasets of various compositions in section 5.3.1, ten training sets and one test set were constructed for each HLA-B7 allele. We trained ANN models for each of the six HLA-B7 molecules (*0702, *3501, *5101, *5301, *5401, and *5501) using the ten datasets to find the composition of training data that results in best predictive performance of the trained system.

A total of 180 simulations were performed for testing: six HLA-B7 alleles, ten training datasets for each allele, with three runs for each dataset. We calculated 540 AROC values (for H, MH, and LMH sets in each simulation). The AROC values ranged from 0.62 to 0.96 with average of 0.79. The maximum standard deviation for repeated runs was 0.086 and average standard deviation was 0.074. The test results were therefore highly reproducible. Sixty-three values (11.7% of the total) represented models that had $0.62 \leq \text{AROC} < 0.7$ and therefore considered not useful. Out of the 63 low-scoring AROC values, 55 of them pertained to the models of HLA-B*5101 and eight of them to those of *5401. The average AROC values for HLA-B*0702, *3501, *5101, *5301, *5401, and *5501 were 0.79, 0.77, 0.69, 0.82, 0.74, and 0.91, respectively. These results indicate good or acceptable predictive performance for the majority of HLA-B7 alleles, except for a marginal predictive performance of HLA-B*5101 and *5401 models.

Figure 5.5 Performance comparison of HLA-B7 ANN models trained using training datasets of various compositions.

The average AROC values of ANN models trained with the ten training datasets are shown in Figure 5.5. The test results for the selected optimal model (one copy of binders) are shown in Table 5.15. Predictions for HLA-B*5101 and *5401 were of marginal accuracy. All other models showed reasonable predictive performance.

Table 5.15 AROC values for blind testing of HLA-B7 ANN models.

	AROC						Average
	*0702	*3501	*5101	*5301	*5401	*5501	
LMH	0.83	0.79	0.70	0.84	0.75	0.94	0.81
MH	0.82	0.76	0.70	0.80	0.74	0.95	0.80
H	0.83	0.76	0.68	0.80	0.74	0.91	0.79

5.3.4 Blind testing of HLA-A2 HMM models

We did blind tests on five HLA-A2 alleles and built five HMM models, one for each of the alleles. The numbers of peptides in training and testing datasets are shown in Table 5.16. For details on blind tests, refer to section 5.3.1. Different from building ANN models, only binders are used to train HMM models.

The assessment of accuracy of the five HLA-A2 HMM models is shown in Table 5.17. Overall, predictions were of high accuracy for peptides binding to HLA-A*0201, *0204, and *0205 (AROC>0.9). The *0206 binding predictions for all peptides and moderate binders were of high accuracy (AROC>0.8). Predictions for *0202 were of moderate accuracy for all peptides (AROC=0.73), but particularly poor for moderate and high affinity peptides (AROC<0.7).

Table 5.16 Blind testing of HMM models for five HLA-A2 alleles: Number of peptides in the training and testing sets for HMM models.

HLA-A2 alleles	Training data	Testing data	
	Binders	Binders	Non-binders
*0201	224	434	1999
*0202	619	45	25
*0204	641	23	224
*0205	648	16	40
*0206	621	43	37

Table 5.17 AROC values for blind testing of HLA-A2 HMM models.

	AROC					
	*0201	*0202	*0204	*0205	*0206	Average
LMH	0.93	0.73	0.92	0.94	0.88	0.93
MH	0.93	0.67	0.92	0.96	0.85	0.93
H	0.92	0.56	0.9	0.91	0.78	0.92

5.3.5 Blind testing of HLA-A3 HMM models

We did blind tests on seven HLA-A3 alleles and built seven HMM models, one for each of the alleles. The numbers of peptides in training and testing datasets are shown in Table 5.18. Different from building ANN models, only binders are used to train HMM models. For details on blind tests, refer to section 5.3.1.

The assessment of accuracy of the seven HLA-A3 molecule prediction models is shown in Table 5.19. Overall, predictions were of a high accuracy for peptide binding to HLA-A*6801 (AROC>0.9). The *0301, *0302, *1101, and *1102 binding predictions were of high accuracy (AROC>0.8). Predictions performance for *3101 and *3301 were particularly poor. The reason for poor performance of predictions for HLAA*3101 and *3301 could be that the test sets for these two molecule contain some kind of bias, or that the training set does not contain sufficient information to infer binding rules in this case.

Table 5.18 Blind testing of HMM models for seven HLA-A3 alleles: Numbers of peptides in training and testing sets for HMM models.

HLA-A3 alleles	Training data	Testing data	
	Binders	Binders	Non-binders
*0301	573	107	89
*0302	534	146	259
*1101	538	142	223
*1102	538	142	211
*3101	636	44	54
*3301	645	35	62
*6801	621	59	638

Table 5.19 AROC values for blind testing of HLA-A3 HMM models.

AROC								
	*0301	*0302	*1101	*1102	*3101	*3301	*6801	Average
LMH	0.94	0.86	0.91	0.86	0.66	0.58	0.95	0.82
MH	0.89	0.83	0.88	0.82	0.56	0.53	0.95	0.78
H	0.85	0.82	0.87	0.80	0.51	0.55	0.95	0.76

5.3.6 Blind testing of HLA-B7 HMM models

We did blind tests on six HLA-B7 alleles and built six HMM models, one for each of the alleles. The numbers of peptides in training and testing datasets are shown in Table 5.20.

Different from building ANN models, only binders are used to train HMM models. For details on blind tests, refer to section 5.3.1.

The assessment of accuracy of the six HLA-B7 prediction models is shown in Table 5.21. The values ranged from 0.64 to 0.92 with average of 0.79 for prediction of all binders (LMH). Overall, predictions were of a high accuracy for peptide binding to HLA-B*0702 and *5501 (AROC > 0.85). The *3501 and *5301 binding predictions were

of acceptable accuracy (AROC > 0.75). Predictions performance for *5101 and *5401 were particularly poor.

Table 5.20 Blind testing of HMM models for six HLA-B7 alleles: Numbers of peptides in training and testing sets for HMM models.

HLA-B7 alleles	Training data	Testing data	
	Binders	Binders	Non-binders
*0702	957	301	1273
*3501	885	373	779
*5101	1092	166	176
*5301	1119	139	211
*5401	1161	97	188
*5501	1176	82	204

Table 5.21 AROC values for blind testing of HLA-B7 HMM models.

	AROC						Average
	*0702	*3501	*5101	*5301	*5401	*5501	
LMH	0.88	0.81	0.65	0.79	0.66	0.92	0.79
MH	0.88	0.78	0.65	0.75	0.65	0.90	0.77
H	0.88	0.79	0.64	0.75	0.66	0.88	0.77

5.3.7 Blind testing of HLA-A2 SVM models

Blind testing was performed to analyze the performance of SVMs for prediction of promiscuous HLA-A2 binding peptides. The training and testing datasets used were same as those used in blind testing of ANN models (Table 5.10). We trained more than 200 different SVM models for each of the five HLA-A2 alleles. Parameter values used during this process were the following: 13 values of trade-off C in equation (2.35) varied from 0.01 to 20, nine values of polynomial degree d in equation (2.33) varied from 2 to 10, and seven values of g in equation (2.34) varied from 0.001 to 1.

Table 5.22 AROC values for blind testing of HLA-A2 SVM models. The three kernel functions are linear, 2-degree polynomial and Gaussian ($g=0.1$ and $C=0.5$).

HLA-A2 alleles	Kernel		
	Linear	Gaussian	Polynomial
*0201	0.92	0.90	0.92
*0202	0.73	0.81	0.76
*0204	0.90	0.93	0.92
*0205	0.93	0.97	0.97
*0206	0.83	0.85	0.85
Average	0.86	0.89	0.88

Table 5.23 AROC values and Pearson correlation coefficients for blind testing of HLA-A2 support vector regression models. The kernel functions was Gaussian ($g=0.1$ and $C=0.5$).

HLA-A2 alleles	Gaussian Kernel	
	AROC	Pearson correlation coefficient
*0201	0.91	0.59
*0202	0.80	0.44
*0204	0.92	0.64
*0205	0.91	0.74
*0206	0.80	0.58
Average	0.87	0.60

Comparison of predictive performance of the three kernels, linear, Gaussian and polynomial, is shown in Table 5.22. The models that have the best performance (in terms of the highest average blind testing AROC value for molecules within a given supertype) are models with a Gaussian kernel ($g=0.1$ and $C=0.5$). AROC values for models with polynomial kernels are very close to those of models with Gaussian kernels. Models using a linear kernel had the lowest average AROC values. Predictions are of excellent accuracy for peptides binding to HLA-A*0201, *0204, and *0205 ($\text{AROC} \geq 0.90$) and good accuracy for HLA-A*0202 and *0206. Overall, predictions are of high accuracy ($\text{AROC}=0.89$) for HLA-A2 SVM models with Gaussian kernels. The predictive performance of support vector regression models using Gaussian kernel ($g=0.1$ and $C=0.5$) is shown in Table 5.23. By comparing the AROC values of Gaussian kernel in

Table 5.22 and Table 5.23, we found SVM models and support vector regression models produced similar accuracy with SVM models performing slightly better (0.89 vs. 0.87). Thus in the following experiments, SVM models have been used for prediction.

5.3.8 Blind testing of HLA-A3 SVM models

Blind testing was performed to analyze the performance of SVMs for prediction of promiscuous HLA-A3 binding peptides. The training and testing datasets used are the same as those used in blind testing of ANN models (Table 5.12). The assessment of accuracy of HLA-A3 prediction models is shown in Table 5.24. Predictions are of excellent accuracy for peptide binding to HLA-A*0301, *1101, *1102, *3301, and *6801. Predictions are of good accuracy for HLA-A*0302 and *3101. Overall, predictions are of almost excellent accuracy (AROC=0.93) for HLA-A3 SVM models with Gaussian kernel. The models that have the best performance (in terms of the highest average blind testing AROC value for molecules within a given supertype) are models with a Gaussian kernel ($g=0.1$ and $C=2$). AROC values for models with a 2-degree polynomial kernel are very close to those of a Gaussian kernel. Models using linear kernel had the lowest average AROC.

Table 5.24 AROC values for blind testing of HLA-A3 SVM models. The three kernel functions are linear, 2-degree polynomial and Gaussian ($g=0.1$ and $C=2$).

HLA-A3 alleles	Kernel		
	Linear	Gaussian	Polynomial
*0301	0.91	0.93	0.92
*0302	0.84	0.86	0.86
*1101	0.94	0.96	0.93
*1102	0.89	0.96	0.91
*3101	0.82	0.87	0.90
*3301	0.81	0.92	0.96

*6801	0.99	0.98	0.98
Average	0.89	0.93	0.92

5.3.9 Blind testing of HLA-B7 SVM models

Blind testing was performed to analyze the performance of SVMs for prediction of promiscuous HLA-B7 binding peptides. The training and testing datasets used are the same as those used in blind testing of ANN models (Table 5.14). The assessment of accuracy of HLA-B7 prediction models is shown in Table 5.25. Predictions are of good accuracy for peptides binding to HLA-B*0702, *3501, *5301, and *5501 ($\text{AROC} \geq 0.80$). Prediction accuracy for HLA-B*5401 is acceptable, while that of *5101 is marginal. Overall, predictions are of almost excellent accuracy ($\text{AROC} = 0.84$) for HLA-B7 SVM models with a Gaussian kernel and a 2-degree polynomial kernel. The models that have the best performance (in terms of the highest average blind testing AROC value for molecules within a given supertype) are as models with a Gaussian kernel ($g = 0.05$ and $C = 0.5$) and 2-degree polynomial kernel. Models using linear kernel had the lowest average AROC.

Table 5.25 AROC values for blind testing of HLA-B7 SVM models. Three kernel functions are linear, 2-degree polynomial and Gaussian ($g = 0.05$ and $C = 0.5$).

HLA-B7 alleles	Kernel		
	Linear	Gaussian	Polynomial
*0702	0.90	0.90	0.91
*3501	0.80	0.82	0.83
*5101	0.61	0.76	0.68
*5301	0.85	0.88	0.87
*5401	0.77	0.77	0.79
*5501	0.95	0.88	0.93
Average	0.81	0.84	0.84

5.3.10 Performance comparison of ANN, HMM, and SVM models

According to the blind testing results reported in section 5.3.1-5.3.9, SVM models showed the best overall performance among the three methods (Figure 5.6). SVM models improved prediction accuracy from moderate to good for A*0202, from good to excellent for A*1102, from poor to good for A*3101, from poor to excellent for the A*3301 molecule, and from poor to moderate for B*5101. This was partially due to the fact that the numbers of peptides related to these molecules are relatively small, and it has been reported that SVM outperforms other prediction methods on smaller datasets. HMM models showed the best performance in predicting peptides binding to A*0201 and A*0301, ANN showed the best performance in predicting peptides binding to B*5501, and both HMM and ANN performed better than SVM in predicting peptides binding to A*0206.

A)

120

B)

C)

Figure 5.6 Blind testing performance comparison. A) HLA-A2 models; B) HLA-A3 models; C) HLA-B7 models.

Student's t-tests have been used to assess if SVM models outperformed HMM and ANN models significantly. Details on t-test can be found in section 2.5.2. The significance level was set to 0.05 and 1-tail paired t-tests have been applied. The results showed that SVM models performed statistically better than ANN and HMM models in

prediction of promiscuous peptides of HLA-A2, -A3, and -B7 supertypes at significance level of 0.05.

Ten-fold cross-validation was performed using all three methods too. The sensitivities of the models were calculated at three levels: SP values of 0.8, 0.9, and 0.95 (Table 5.26). Among all the models for the three supertypes, SVM models were of the highest sensitivities. Especially when specificity was high (=0.95), the sensitivities of SVM models were markedly higher than those of ANN and HMM models.

Table 5.26 Sensitivities in 10-fold cross-validation of SVM, ANN, and HMM models. The values are given for three levels of specificity, 0.8, 0.9, and 0.95. A) HLA-A2 supertype (Dataset1); B) HLA-A3 supertype (Dataset2); C) HLA-B7 supertype (Dataset3).

A)

Specificity	Sensitivity		
	SVM	ANN	HMM
0.80	0.96	0.95	0.69
0.90	0.90	0.84	0.55
0.95	0.76	0.55	0.42

B)

Specificity	Sensitivity		
	SVM	ANN	HMM
0.80	0.97	0.86	0.56
0.90	0.92	0.66	0.36
0.95	0.84	0.41	0.24

C)

Specificity	Sensitivity		
	SVM	ANN	HMM
0.80	0.90	0.85	0.56
0.90	0.74	0.64	0.41
0.95	0.55	0.38	0.29

Unlike prediction models developed for blind testing in section 5.3.1-5.3.9, the final prediction models used in MULTIPRED1 were trained using all available data in Dataset1, Dataset2, and Dataset3. Thus, it is reasonable to expect that test results

presented in section 5.3.1-5.3.9 represent an underestimate of the performance of the models trained using complete datasets.

Stratified 10-fold cross-validation was performed on peptides related to A*0201 and A*0302. In stratified 10-fold cross-validation, a dataset is randomly divided into 10 sets with approximately equal size and class distributions. And the peptides in the training data that were similar (only one amino acid different) to any peptide from the test set were removed. The AROC values obtained from 10-fold cross-validations for HLA-A*0201 and HLA-A*0302 are shown in Figure 5.7. In both cases, the AROC values of 10-fold cross-validations were higher than those of blind tests by 0.05. This is in accordance with our assumption that blind testing accuracy is in fact an underestimate of the actual accuracy of the method.

Figure 5.7 AROC values of SVM models in blind tests and 10-fold cross-validations for predicting of peptides binding to HLA-A*0201 and *0302.

5.3.11 Validation of HLA-A2 prediction models using experimental results

Validation of the HLA-A2 models was conducted using experimental data from HPV type 16 E6 (NCBI Accession: P03126) and E7 (NCBI Accession: P03129) proteins [214] and a study on human cytomegalovirus (HCMV) antigens [215, 216]. HPV infection is associated with the development of cervical cancers. As the most prevalent HPV type worldwide, HPV type 16 is present in more than 99% of cervical cancers [217]. Most healthy adults who acquire primary HCMV infection do not have long-term health problems. However around 80% of infants who acquire congenital HCMV infection display a variable pattern of pathological sequelae within the first few years of life that may include hearing loss, vision impairment, and mental retardation [215]. Cell-mediated immunity is essential for control of HPV and HCMV infections.

Dataset1 was checked for duplicate 9-mer peptides pertaining to HPV E6 and E7 proteins. 23 duplicate 9-mer peptides were removed from Dataset1. The resulting training datasets for HLA-A2 contained 3027 9-mer peptides (651 binders and 2376 non-binders). The two most immunologically dominant HCMV antigens, pp65 (NCBI Accession: CAA35357) and IE-1 (NCBI Accession: P13202), which contain HLA-A2 restricted epitopes, were used for validation of our prediction models. These two proteins do not have overlap with Dataset1.

Experimentally identified HLA-A*0201 binding peptides in HPV E6 and E7 are listed in Table 5.27 [214, 218]. Peptides from HCMV antigens were pre-selected using SYFPEITHI [22] and BIMAS [97], then tested to identify T-cell epitopes using ELISPOT assays (Table 5.28) [215].

Table 5.27 HLA-A*0201 binding peptides in HPV type 16 E6 and E7 proteins. “Pos.”, “H”, and “L” stand for the start position of a peptide, high binding affinity, and low binding affinity.

HLA-A2 binding peptides in HPV E6			HLA-A2 binding peptides in HPV E7		
Pos.	Peptides	Binding affinity	Pos.	Peptides	Binding affinity
7	AMFQDPQER	L	7	TLHEYMLDL	H
18	KLPQLCTEL	H	11	YMLDLQPET	H
26	LQTTIHDI	L	12	MLDLQPETT	H
52	FAFRDLCIV	H	66	RLCVQSTHV	L
			82	LLMGTLGIV	H
			85	GTLGIVCPI	H
			86	TLGIVCPIC	H

Table 5.28 HLA-A2 restricted T-cell epitopes from HCMV pp65 and IE-1. The peptides highlighted in yellow are commonly recognized T-cell epitopes. “Pos.” stands for the start position of a T-cell epitope.

HLA-A2-restricted T-cell epitopes in pp65		HLA-A2-restricted T-cell epitopes in IE-1	
Pos.	Peptides	Pos.	Peptides
40	RLLQTGIHV	81	VLAELVKQI
120	MLNIPSINV	125	CLQNALDIL
320	LMNGQQIFL	303	SLLSEFCRV
491	ILARNLVPM	316	VLEETSVML
495	NLVPMVATV		
522	RIFAELEGV		

Table 5.29 The top 10% of predicted HLA-A*0201 9-mer binders (15 peptides) in HPV E6 in descending order of binding scores. “Pos” stands for the start position of a 9-mer peptide. The bold ones are the true positive predictions. All models picked up the two high binders, but none of them picked up the two weak binders.

ANN		HMM		SVM	
Pos.	Peptide	Pos.	Peptide	Pos.	Peptide
52	FAFRDLCIV	25	ELQTTIHDI	18	KLPQLCTEL
18	KLPQLCTEL	18	KLPQLCTEL	52	FAFRDLCIV
105	DLLIRCINC	29	TIHDIILEC	106	LLIRCINCQ
102	PLCDLLIRC	21	QLCTELQTT	89	SLYGTTLQ
106	LLIRCINCQ	72	KCLKFYSKI	25	ELQTTIHDI
30	IHDIILECV	22	LCTELQTTI	30	IHDIILECV
89	SLYGTTLQ	68	AVCDKCLKF	29	TIHDIILEC
25	ELQTTIHDI	1	MHQKRTAMF	27	QTTIHDIIL
43	QLLRREVYD	79	KISEYRHYC	43	QLLRREVYD
27	QTTIHDIIL	75	KFYISKISEY	105	DLLIRCINC
37	CVYCKQQLL	44	LLRREVYDF	102	PLCDLLIRC
125	HLDKKQRFH	127	DKKQRFHNI	37	CVYCKQQLL
51	DFAFRDLCI	89	SLYGTTLQ	125	HLDKKQRFH
29	TIHDIILEC	94	TLEQQYNKP	103	LCDLLIRCI

E6 protein is 158 amino acid long and was decomposed into 150 overlapping 9-mer peptides. The top 10% of predictions are thus 15 peptides. In Table 5.29, the top 10% of predicted HLA-A*0201 9-mer binders in HPV E6 are listed in descending order of binding scores. The peptides highlighted in bold are the experimentally validated binders (true positive predictions). All three models picked up the two high binders, E6-18 and E6-52, and missed out the two weak binders E6-7 and E6-26.

E7 protein is 98 amino acid long. It can be decomposed into 90 overlapping 9-mer peptides. The top 10% of predicted HLA-A*0201 binders in E7 protein (9 peptides) are shown in Table 5.30. The peptides highlighted in bold are the true positive predictions. ANN models picked up all seven binders, SVM models picked up five binders and missed out two binders, while HMM models picked up four binders and missed out three binders.

Table 5.30 The top 10% of predicted HLA-A*0201 9-mer binders (nine peptides) in HPV E7 in descending order of binding scores. “Pos” stands for the start position of a 9-mer peptide. The highlighted ones are the true positive predictions. ANN models picked up all seven binders.

ANN		HMM		SVM	
Pos.	Peptide	Pos.	Peptide	Pos.	Peptide
82	LLMGTLGIV	7	TLHEYMLDL	82	LLMGTLGIV
7	TLHEYMLDL	68	CVQSTHVDI	7	TLHEYMLDL
81	DLLMGTLGI	66	RLCVQSTHV	66	RLCVQSTHV
86	TLGIVCPIC	82	LLMGTLGIV	81	DLLMGTLGI
66	RLCVQSTHV	53	NIVTFCKC	85	GTLGIVCPI
85	GTLGIVCPI	78	TLEDLLMGT	86	TLGIVCPIC
11	YMLDLQPET	14	DLQPETTDL	51	HYNIVTFCC
14	DLQPETTDL	63	STLRLCVQS	64	TLRLCVQST
12	MLDLQPETT	11	YMLDLQPET	14	DLQPETTDL

HCMV pp65 protein is 561 amino acid long and was decomposed into 553

overlapping 9-mer peptides. The top 5% of predictions are thus 28 peptides. The top 5% of predictions of HLA-A2 binders in HCMV pp65 (Table 5.31) by each of ANN, HMM, and SVM models covered all six identified HLA-A2 restricted T-cell epitopes in pp65 protein listed in Table 5.28. Since it is a selective study in [215], we cannot tell whether the rest 22 predicted peptides are false positive or true positive.

HCMV IE-1 protein is 491 amino acid long and was decomposed into 483 overlapping 9-mer peptides. The top 5% of predictions are thus 24 peptides. The top 5% of predictions of HLA-A2 binders in HCMV IE-1 (Table 5.32) by ANN and SVM covered all four HLA-A2 restricted T-cell epitopes in IE-1 protein listed in Table 5.28, while HMM predictions missed out one epitope, 125-133.

Table 5.31 The top 5% of predicted HLA-A2 binders in HCMV pp65 (28 entries) in descending order of binding scores. “Pos” stands for the start position of a 9-mer peptide. The peptides highlighted in bold are recognized T-cell epitopes. All three models picked up all T-cell epitopes in pp65.

ANN		HMM		SVM	
Pos	Peptide	Pos	Peptide	Pos	Peptide
120	MLNIPSINV	155	QMWQARLTV	40	RLLQTGIHV
495	NLVPMVATV	495	NLVPMVATV	120	MLNIPSINV
40	RLLQTGIHV	286	KISHIMLDV	495	NLVPMVATV
312	GLSISGNLL	86	EVENVSVNV	347	ALFFFDIDL
320	LMNGQQIFL	218	VIGDQYVKV	14	VLGPISGHV
14	VLGPISGHV	40	RLLQTGIHV	312	GLSISGNLL
347	ALFFFDIDL	331	QAIRETVEL	320	LMNGQQIFL
491	ILARNLVPM	210	NTRATKMVQ	286	KISHIMLDV
318	NLLMNGQQI	329	EVQAIRETV	115	ALPLKMLNI
155	QMWQARLTV	491	ILARNLVPM	227	YLESFCEDV
108	PMSIYVYAL	227	YLESFCEDV	349	FFFDIDLLL
296	FTSHEHFGL	200	CAHELVCSM	274	VLCPKNMII
194	ALRHVVCAH	325	QIFLEVQAI	239	KLFMHVTLG
10	EMISVLGPI	519	DIYRIFAEL	522	RIFAELEGV
274	VLCPKNMII	194	ALRHVVCAH	108	PMSIYVYAL
239	KLFMHVTLG	120	MLNIPSINV	54	SLILVSQYT
34	VLPHETRL	412	TTERKTPRV	141	HLPVADAVI
110	SIYVYALPL	82	FTGSEVENV	82	FTGSEVENV

127					
54	SLILVSQYT	320	LMNGQQIFL	296	FTSHEHFGL
522	RIFAELEGV	21	HVLKAVFSR	318	NLLMNGQQI
286	KISHIMLDV	274	VLCPKNMII	10	EMISVLGPI
141	HLPVADAVI	211	TRATKMQUI	491	ILARNLVPM
279	NMIKPGKI	522	RIFAELEGV	34	VLPHETRL
227	YLESFCEDV	290	IMLDVAFTS	155	QMWQARLTV
290	IMLDVAFTS	197	HVVCAHELV	442	SATACTSGV
519	DIYRIFAEL	425	AMAGASTSA	110	SIYVYALPL
115	ALPLKMLNI	409	ELVTTERKT	216	MQVIGDQYV
349	FFFDIDLLL	148	VIHASGKQM	194	ALRHVVCAH

Table 5.32 The top 5% of predicted HLA-A2 binders in HMCV IE-1 (24 entries) in descending order of binding scores. “Pos” stands for the start position of a 9-mer peptide. The peptides highlighted in bold are recognized T-cell epitopes. ANN and SVM models picked up all T-cell epitopes in IE-1, while HMM model missed out one epitope.

ANN		HMM		SVM	
Pos	Peptide	Pos	Peptide	Pos	Peptide
303	SLLSEFCRV	81	VLAELVKQI	303	SLLSEFCRV
81	VLAELVKQI	342	RIEEICMKV	272	FMDILTTCV
272	FMDILTTCV	85	LVKQIKVRV	355	ILGADPLRV
125	CLQNALDIL	338	VMKRRIEEI	81	VLAELVKQI
316	VLEETSVML	248	AYAQKIFKI	125	CLQNALDIL
355	ILGADPLRV	245	EIMAYAQKI	304	LLSEFCRVL
304	LLSEFCRVL	103	HMLKKYTQT	297	TMYGGISLL
129	ALDILDKVH	316	VLEETSVML	316	VLEETSVML
297	TMYGGISLL	199	ELRRKMMYM	360	PLRVCSPSV
256	ILDEERDKV	303	SLLSEFCRV	311	VLCCYVLEE
49	SLGDPLFPE	169	ACIKELHDV	323	MLAKRPLIT
173	ELHDVSKGA	275	ILTTCVETM	393	GVSSSDSLV
323	MLAKRPLIT	263	KVLTHIDHI	342	RIEEICMKV
330	ITKPEVISV	230	QAMAALQNL	256	ILDEERDKV
263	KVLTHIDHI	330	ITKPEVISV	413	PLSSVIVAE
128	NALDILDKV	135	KVHEPFEEEM	128	NALDILDKV
381	EEAIVAYTL	315	YVLEETSVM	169	ACIKELHDV
84	ELVKQIKVR	57	ELAEESLKT	248	AYAQKIFKI
360	PLRVCSPSV	346	ICMKVFAQY	49	SLGDPLFPE
311	VLCCYVLEE	256	ILDEERDKV	296	MTMYGGISL
322	VMLAKRPLI	61	ESLKTFEQV	330	ITKPEVISV
152	SMYENYIVP	267	HIDHIFMDI	263	KVLTHIDHI
57	ELAEESLKT	39	MLRKEVNSQ	411	TIPLSSVIV
39	MLRKEVNSQ	31	KATTFLQTM	184	KLGGALQAK

5.3.12 Validation of HLA-A3 prediction models using experimental results

Validation of the HLA-A3 models was conducted using experimental data from HPV type 16 E6 and E7 proteins [214] and a study on HCMV antigens [215, 216]. Dataset2 was checked for the duplicate 9-mer peptides pertaining to HPV E6 and E7 proteins and 52 duplicate 9-mer peptides were removed from Dataset2. The resulting training datasets contained 2164 9-mer peptides (653 binders and 1511 non-binders). Experimentally identified HLA-A*0301 binding peptides in HPV E6 and E7 are listed in Table 5.33 [214, 218].

Table 5.33 HLA-A*0301 binding peptides in HPV E6 and E7 proteins. “Pos.”, “H”, and “L” stand for the start position of a peptide, high binding affinity, and low binding affinity.

HLA-A3 binding peptides in HPV E6			HLA-A3 binding peptides in HPV E7		
Pos.	Peptides	Binding affinity	Pos.	Peptides	Binding affinity
7	AMFQDPQER	H	89	IVCPICSQK	L
33	IILECVYCK	L			
42	QQLLRREVY	L			
59	IVYRDGNPY	H			
75	KFYISKISEY	L			
89	SLYGTTLQ	L			
93	TTLEQQYNK	H			
125	HLDDKKQRFH	H			
143	AMSAARSSR	L			

Table 5.34 The top 10% of predicted HLA-A*0301 9-mer binders (15 peptides) in HPV E6 in descending order of binding scores. “Pos.” stands for the start position of a 9-mer peptide. The bold ones are the true positive predictions. HMM and SVM models picked up six out of nine binders.

ANN		HMM		SVM	
Pos.	Peptide	Pos.	Peptide	Pos.	Peptide
33	IILECVYCK	93	TTLEQQYNK	107	LIRCINCQK
145	SCCRSSRTR	7	AMFQDPQER	143	CMSCCRSSR
68	AVCDKCLKF	39	YCKQQLLR	75	KFYISKISEY
143	CMSCCRSSR	107	LIRCINCQK	134	NIRGRWTGR
107	LIRCINCQK	120	EEKQRHLDK	59	IVYRDGNPY
7	AMFQDPQER	38	VYCKQQLR	89	SLYGTTLQ
39	YCKQQLLR	33	IILECVYCK	33	IILECVYCK

129					
146	CCRSSRTRR	145	SCCRSSRTR	79	KISEYRHYC
93	TTLEQQYNK	89	SLYGTTLQ	7	AMFQDPQER
128	KKQRFHNIR	146	CCRSSRTRR	71	DKCLKFYSK
76	FYSKISEYR	116	PLCPEEKQR	97	QQYNKPLCD
64	GNPYAVCDK	75	KFYKISEY	131	RFHNIRGRW
101	KPLCDLLIR	143	CMSCCRSSR	68	AVCDKCLKF
67	YAVCDKCLK	134	NIRGRWTGR	145	SCCRSSRTR
81	SEYRHYCYS	48	EVYDFAFRD	146	CCRSSRTRR

The top 10% of predicted HLA-A3 binders (15 peptides) in E6 protein are shown in Table 5.34. Each of HMM and SVM models picked up six binders, while ANN models picked up four out of the nine identified HLA-A*0301 binding peptides in HPV type 16 E6 listed in Table 5.33.

The top 10% of predicted HLA-A3 binders (nine peptides) in E7 protein are shown in Table 5.35. The peptides highlighted in bold are the true positive predictions. All models picked up E7-89, the experimentally identified HLA-A3 binder.

Table 5.35 The top 10% of predicted HLA-A*0301 9-mer binders (nine peptides) in HPV E7 in descending order of binding scores. “Pos” stands for the start position of a 9-mer peptide. The bold ones are the true positive predictions. All models picked up the binder.

ANN		HMM		SVM	
Pos.	Peptide	Pos.	Peptide	Pos.	Peptide
89	IVCPICSQK	27	QLNDSSEEE	89	IVCPICSQK
58	CCKCDSTLR	89	IVCPICSQK	21	DLYCYEQLN
85	GTLGIVCPI	58	CCKCDSTLR	7	TLHEYMLDL
50	AHYNIVTFC	73	HVDIRTLED	44	QAEPDRAHY
55	VTFCKCDS	44	QAEPDRAHY	37	EIDGPAGQA
52	YNIVTFCKK	66	RLCVQSTHV	66	RLCVQSTHV
45	AEPDRAHYN	68	CVQSTHVDI	53	NIVTFCKKC
88	GIVCPICSQ	70	QSTHVDIRT	88	GIVCPICSQ
63	STLRLCVQS	7	TLHEYMLDL	27	QLNDSSEEE

5.3.13 Validation of HLA-B7 prediction models using experimental results

Validation of the HLA-B7 models was conducted using experimental data from a selective study on Telomerase reverse transcriptase (TRT) (NCBI Accession: O14746).

Dataset3 was checked for duplicate 9-mer peptides pertaining to TRT protein. There is no overlapping 9-mer peptide between them. Peptides from TRT antigens were pre-selected using SYFPEITHI [22] and BIMAS [97], then tested using in vivo immunization of HLA-B7 transgenic mice, binding stabilization assay on T2-B7 cells by flow cytometry and a competitive solid-phase radioimmunoassay on immobilized purified HLA-B7 molecule [219]. There are five HLA-B7 9-mer binders, TRT-277, TRT-342, TRT-464, TRT-1107, and TRT-1123. TRT-1123 has measurable binding affinity to B*0702, B*3501, B*5101 and B*5301. TRT-1107 and TRT-342 have measurable binding affinity to B*0702, B*3501 and B*5301. TRT-277 has measurable binding affinity to B*0702 and B*3501.

Table 5.36 The top 2% of predicted HLA-B7 binders (22 peptides) in TRT in descending order of binding scores. “Pos” stands for the start position of a 9-mer peptide. The bold ones are the true positive predictions. ANN models picked up all five binders.

ANN		HMM		SVM	
Pos.	Peptide	Pos.	Peptide	Pos.	Peptide
922	MPAHGLFPW	277	RPAEEATSL	1107	LPGTTLTAL
1107	LPGTTLTAL	123	LPNTVTDAL	1123	LPSEDFKTIL
1123	LPSEDFKTIL	720	IPQDRLTEV	922	MPAHGLFPW
1028	NPTFFLRVI	243	APEPERTPV	1	MPRAPRCRA
123	LPNTVTDAL	1	MPRAPRCRA	672	RPGLLGASV
1	MPRAPRCRA	1123	LPSEDFKTIL	43	DPAAFRALV
342	RPSFLLSSL	379	LPRLPQRYW	1065	LPSEAVQWL
672	RPGLLGASV	1065	LPSEAVQWL	123	LPNTVTDAL
153	LARCALFVL	274	SPARPAEEA	1028	NPTFFLRVI
43	DPAAFRALV	1107	LPGTTLTAL	613	RPALLTSRL
720	IPQDRLTEV	922	MPAHGLFPW	928	FPWCGLLLD
1065	LPSEAVQWL	444	DPRRLVQLL	22	LPLATFVRR
91	RGAKNVLAF	234	LPKRPRRGA	63	RPPPAAPSF
369	RPWMPGTPR	784	SPLRDAVVI	58	VPWDARPPP
613	RPALLTSRL	63	RPPPAAPSF	342	RPSFLLSSL
277	RPAEEATSL	812	FMCHHAVRI	626	KPDGLRPIV
784	SPLRDAVVI	613	RPALLTSRL	720	IPQDRLTEV
58	VPWDARPPP	631	RPIVNMDYV	372	MPGTPRRLP
416	RAAVTPAAG	770	QPYMRQFVA	420	TPAAGVCAR
63	RPPPAAPSF	68	APSRQVSC	444	DPRRLVQLL

131

379	LPRLPQRYW	874	TPHLTHAKT	907	FPVEDEALG
464	FVRACLRL	43	DPAAFRALV	315	RPWDTPCPP

TRT protein is 1032 amino acid long and was decomposed into 1124 overlapping 9-mer peptides. The top 2% of predictions are thus 22 peptides. The top 2% of predictions of HLA-B7 binders in TRT (Table 5.36) by ANN covered all five experimental HLA-B7 binders, while both SVM and HMM models managed to pick up three binders. All the models picked up the two most promiscuous binders, TRT-1123 and TRT-1107.

5.3.14 Performance comparison of MULTIPRED1 models with other prediction systems

Several web-based systems have been developed and widely used for prediction of HLA binders, such as SYFPEITHI based on binding motifs [22], BIMAS based on quantitative matrices [97], SMM based an improved matrix-based algorithm called stabilized matrix method [102], RANKPEP based on Position Specific Scoring Matrices [100], and SVMHC based on SVM [115]. To assess the performance of our models, the results of our models in predicting individual 9-mer binders was compared with the prediction tools mentioned above.

A large experimental dataset of quantitative MHC–peptide binding data was recently made available [178]. Using this dataset, Peters *et al.* compared the performance of several bioinformatics approaches, including MULTIPRED ANN and HMM models, in predicting MHC binding peptides. The performance of these models has been reported in Tables 2 and 3 of the paper. We used the same dataset for evaluation of our HLA-A2/A3 SVM models. In the Peters dataset, the IC50 value used to separate binders and non-binders is 500 nM, whereas in Dataset1 and Dataset2, the IC50 value used to separate

binders and non-binders is 5000 nM. The difference in the threshold setting results in some discrepancy between our datasets and the Peters dataset. Some identical peptides between Dataset1, Dataset2, and the Peters dataset have been classified into different classes. The binding affinities of such peptides in Dataset1 and Dataset2 were modified to the binding affinities in the Peters dataset and these overlapping peptides were removed from the test sets. The number of 9-mer peptides in the Peters dataset related to HLA-A2/A3 alleles, the number of overlapping 9-mer peptides between Dataset1/Dataset2 and the Peters dataset and the number of non-overlapping peptides used for testing of the HLA-A2/A3 SVM models are shown in

Table 5.37 and Table 5.38, respectively. The testing datasets are available for download at <http://antigen.i2r.a-star.edu.sg/multipred1/data>.

Table 5.37 Number of 9-mer peptides in the Peters dataset related to HLA-A2 alleles; number of overlapping peptides between Dataset1 and the Peters dataset; number of non-overlapping peptides (binders/non-binders) used for testing of the HLA-A2 SVM model.

	A*0201	A*0202	A*0203	A*0206	A*6802	A*6901
The Peters Dataset	3089	1447	1443	1437	1434	833
Overlapping	240	54	48	47	47	0
Non-overlapping	2849	1393	1395	1390	1387	833
(binders/non-binders)	1024/1825	611/782	600/795	480/910	387/1000	86/747

Table 5.38 Number of 9-mer peptides in the Peters dataset related to HLA-A3 alleles; number of overlapping peptides between Dataset2 and the Peters dataset; number of non-overlapping peptides (binders/non-binders) used for testing of the HLA-A3 SVM model.

	A*0301	A*1101	A*3101	A*3301	A*6801
The Peters Dataset	2094	1985	1869	1140	1141
Overlapping	97	102	71	70	69
Non-overlapping	1997	1883	1798	1070	1072
	452/1545	618/1265	399/1399	161/909	455/617

Table 5.39 AROC values of HLA-A2 SVM model for predictions of binding peptides for A*0201, A*0202, A*0203, A*0206, A*6802 and A*6901 using the Peters dataset. AROC values of ANN and HMM models are taken from [178]. MULTIPRED ANN model had the highest AROC values in predicting peptides binding to A*0202, *0203, and *0206 among the external tools evaluated.

	A*0201	A*0202	A*0203	A*0206	A*6802	A*6901
SVM AROC	0.91	0.83	0.82	0.79	0.74	0.81
ANN AROC	0.883	0.793	0.788	0.735	<0.643	-
HMM AROC	0.796	<0.793	<0.788	<0.735	<0.643	-

Table 5.40 AROC values of HLA-A3 SVM model for prediction of binding peptides for A*0301, A*1101, A*3101, A*3301 and A*6801 using the Peters dataset. AROC values of ANN and HMM models are taken from [178]. MULTIPRED ANN model had the highest AROC values in predicting peptides binding to A*0301 and *1101 among the external tools evaluated.

	A*0301	A*1101	A*3101	A*3301	A*6801
SVM AROC	0.87	0.89	0.83	0.79	0.76
ANN AROC	0.851	0.869	<0.829	<0.807	<0.77
HMM AROC	<0.851	0.832	<0.829	<0.807	<0.77

MULTIPRED ANN models showed best performance among all the external tools evaluated in [178] in predicting 9-mer peptides binding to A*0202, A*0203, A*0206, A*0301, and A*1101. The AROC values of ANN models in predicting these alleles are highlighted in bold in Table 5.39 and Table 5.40. As the prediction tools developed by Peters were trained using much larger datasets, we only focus on comparing the performance of our models with other external tools evaluated in the paper. The AROC values of HLA-A2 SVM model for predictions of binding peptides for HLA-A2 supertype alleles (A*0201, A*0202, A*0203, A*0206, A*6802, and A*6901) using the Peters dataset show marked improvement compared to ANN and HMM models for all studied alleles (Table 5.39). We note that HLA-A2 SVM model performs well in predicting A*6901 binding peptides (AROC=0.81) although only four peptides related to A*6901 in Dataset1 were used for model training. Except for A*0201, the SVM model showed best performance in predicting peptides binding to four HLA-A2 alleles (A*0202,

A*0203, A*0206, and A*6802) out of the five alleles evaluated among all the external tools evaluated in [178]. For prediction of 9-mer peptides binding to A*0201, SMM performed best with AROC value being 0.922. AROC values of the HLA-A3 SVM model are equal or better for three of five studied HLA-A3 alleles (A*0301, A*1101, and A*3101) than the best external tool (Table 5.40). We note that only five A*3301 related peptides were in Dataset2 for training of HLA-A3 SVM model.

In Dataset2, there were 697 peptides related to A*6801 with 8.5% of them being binders and 91.5% being non-binders. To understand why HLA-A3 SVM model performed poorly in predicting peptide binding to A*6801 (Table 5.40), we performed additional analysis at the peptide sequence level. According to the SYFPEITHI A*6801 binding motif, anchor positions are position two and nine, with T and V being favored at position two while R and K at position nine. Figure 5.8A shows the amino acid frequency in each position of A*6801 binders in Dataset2. The green and red bars represent the amino acid frequencies in position two and nine, respectively. All binders have either R or K at position nine, with 55.10% of them having R and 44.90% having K. At position two, beside the two favorable amino acids T (frequency 14.29%) and V (frequency 26.53%), I and L also appear often at frequencies 26.53% and 16.33%. For non-binders in Dataset2, the frequencies of R and K are 7.87% and 8.33% at position nine, the frequencies of T, V, I, and L at position two are 6.64%, 3.86%, 9.26%, and 9.88% (Figure 5.8B). Comparing Figure 5.8A and Figure 5.8B, there are obvious differences in frequencies of R and K at position nine and V at position two.

Figure 5.8C shows the amino acid frequency in each position of the true positive predictions using the Peters dataset. The prediction threshold used to distinguish binders

from non-binders was set to the score where the prediction specificity is 80%. The amino acid distribution in position two and nine is very similar to that in Figure 5.8A. At position nine, the frequencies of R and K are 58.49% and 39.63%. At position two, the frequencies of T, V, I, and L are 19.25%, 20.75%, 17.74%, and 15.09%.

Figure 5.8D shows the amino acid frequency in each position of the false positive predictions using the Peters dataset. The amino acid distribution in position two and nine is also similar to that in Figure 5.8A and C. At position nine, the frequencies of R and K are 70.90% and 25.37%. At position two, the frequencies of T, V, I, and L are 5.97%, 14.18%, 11.19%, and 37.31%. Some non-binders might be misclassified as binders due to the similar amino acid distribution as that of binders rather than non-binders in training dataset.

Figure 5.8E shows the amino acid frequency in each position of the false negative predictions in the Peters dataset. At position nine, besides R and K, the frequency of Y is also quite high (28%). At position two, the frequencies of T, V, I, and L are 17.37%, 3.16%, 14.21%, and 11.58%, which are quite similar to those of non-binders. Perhaps some binders were misclassified as non-binders because their amino acid distributions were quite different from that of binders in Dataset2, but similar to that of non-binders at one of the anchor positions, position two.

From the comparison, we found the SVM model was able to classify peptides based on the knowledge learnt from the training data. The poor performance of the SVM model in predicting peptides binding to A*6801 was because the limited number of binders in the training dataset do not represent the diverse pattern of A*6801 binding peptides. This emphasizes the need for use of representative datasets for development of MHC binding

prediction servers. While the accuracy of predictions will improve by the addition of new data to the training sets, it is important to note that about a quarter of MHC class I binding peptides lack canonical binding motifs (A. Sette, personal communication). These peptides are underrepresented in the current datasets, which are used both for training and testing, and therefore the values of accuracy of predictive methods are likely to be somewhat lower than reported, across all comparative studies.

A)

B)

137

C)

D)

138

E)

Figure 5.8 Amino acid frequency in each position of 9-mer peptides. A) Peptides binding to A*6801 in Dataset2; B) Peptides non-binding to A*6801 in Dataset2; C) True positive predictions; D) False positive predictions; E) False negative predictions.

5.4 Discussion

In blind tests on HLA-A2 ANN models, the ANN models predicted peptide binding to HLA-A*0205 with excellent accuracy (AROC>0.90), to HLA-A*0201, HLA-A*0204, and HLA-A*0206 with very good accuracy (AROC>0.85) (Table 5.11). The accuracy of predicting peptide binding to HLA-A*0202 was marginal. HLA-A*0202 has peptide contact environment for a major anchor position, position two, identical to that of HLA-A*0201 (Table 5.4 and Table 5.7). Its contact environment for the middle of the peptide and for an anchor position, position nine, is identical to that of HLA-A*0205 (Table 5.4 and Table 5.7). It is possible that the similarity to both HLA-A*0201 and HLA-A*0205 made training difficult. Another possible reason is that the number of peptides related to HLA-A*0205 (56 peptides) is much smaller than that of HLA-A*0201 (2439 peptides) in Dataset1. The unbalanced dataset may have distorted predictions.

In blind tests on HLA-A3 ANN models, except for a marginal predictive performance of HLA-A*3101 and *3301 models, the models show good or excellent predictive performance for the majority of HLA-A3 alleles (AROC>0.8) (Table 5.13). One possible reason for poor performance of HLA-A*3101 and *3301 models may be the small number of data points in the testing datasets (both testing datasets contain less than 100 data points).

In blind tests on HLA-B7 ANN models, HLA-B*0702, *5301, and *5501 models show good or excellent predictive performance (AROC>0.8) (Table 5.15). As shown in Table 5.3, the largest number of peptides in Dataset3 are related to HLA-B*0702 (1574 peptides). The non-conserved contact residues of HLA-B*3501 are very different from those of *0702, 12 out of 18 are different (Table 5.9). The marginal performance of HLA-B*3501 models in blind tests may be due to its different contact residue environment from that of *0702 and the fact that peptides related to *0702 comprise the majority of the training dataset for HLA-B*3501 models. Out of the 18 non-conserved contact residues of HLA-B*5101, 15 are different from those of *0702. Out of the 18 non-conserved contact residues of allele HLA-B*5401, 12 are different from those of *0702. The poor performance of HLA-B*5101 and *5401 models in the blind test may be also due to their different contact residue environment from that of *0702.

In blind tests on HLA-A2 HMM models, the models predicted peptide binding to HLA-A*0201, *0204, and *0205 with excellent accuracy (AROC>0.90). Compared to ANN models, HLA-A*0201 and *0204 HMM models perform better, while HLA-A*0202 and *0206 ANN models perform better than the corresponding HMM models (Table 5.11 and Table 5.17).

In blind tests on HLA-A3 HMM models, HLA-A*0301 and *3301 HMM models did not perform as well as the corresponding ANN models (Table 5.13 and Table 5.19). Both ANN and HMM HLA-A*6801 models have excellent accuracy ($\text{AROC} \geq 0.95$). HLA-A*1101 and *1102 ANN and HMM models have similar accuracy.

In blind tests on HLA-B7 HMM models, the HMM models predicted peptide binding to HLA-B*0702 and *5501 with good accuracy ($\text{AROC} > 0.85$). Compared to HLA-B7 ANN models, HLA-B*0702 and *3501 HMM models perform better, while HLA-B*5101, *5301, *5401, and *5501 ANN models perform better than the corresponding HMM models (Table 5.15 and Table 5.21).

In blind tests of SVM models, SVM models improved accuracy for predicting peptide binding to A*0202 from moderate to good, to A*1102 from good to excellent, to A*3101 from poor to good, to A*3301 from poor to excellent, to B*5101 from poor to moderate (Figure 5.6). Student t-test results showed that SVM models performed statistically better than ANN and HMM models in prediction of promiscuous peptides of HLA-A2, -A3, and -B7 supertypes at a significance level of 0.05.

Unlike prediction models developed for blind testing in section 5.3.1-5.3.9, the final prediction models used in MULTIPRED1 were trained using all available data in Dataset1, Dataset2, and Dataset3. Thus, it is reasonable to expect that test results presented in section 5.3.1-5.3.9 represent an underestimate of the performance of the models trained using complete datasets. Comparison of AROC values for SVM models in blind testing and 10-fold cross-validation for HLA-A*0201 and *0302 (Figure 5.7) showed that the AROC values of 10-fold cross-validations were higher than those of blind tests by 0.05. This is in accordance with our assumption that blind testing accuracy

is in fact an underestimate of the actual accuracy of the method.

In 10-fold cross validation, the accuracy of SVM models for all three supertypes was excellent ($\text{AROC} \geq 0.90$), the accuracy of all ANN models was good ($\text{AROC} \geq 0.80$), the accuracy of HLA-A2 and A3 HMM models was good too ($\text{AROC} \geq 0.80$), however the AROC value of HLA-B7 HMM models was 0.79.

When validated using experimentally identified 9-mer binders in HPV type 16 E6 and E7 proteins and HCMV antigens (pp65 and IE-1), the HLA-A2 ANN, HMM, and SVM models managed to pick up 19, 15, and 17 binders from the 21 experimental HLA-A2 binders, respectively. The HLA-A3 ANN, HMM, and SVM models managed to pick up five, seven, and seven binders from the ten experimental HLA-A3 binders, respectively. The validation results of ANN, HMM, and SVM models suggest that they are reasonably accurate. The prediction performance will be further improved by the increased number of training data and additional rigorous testing strategies.

Performance of ANN and models was compared with other publicly available prediction systems by Peters *et al.* ANN models showed best performance among all the external tools evaluated in [178] in predicting 9-mer peptides binding to A*0202, A*0203, A*0206, A*0301, and A*1101. We tested SVM model using the same dataset and found that SVM show best performance in predicting 9-mer peptides binding to A*0202, A*0203, A*0206, A*6802, A*0301, A*1101, and A*3101.

Although MULTIPRED1 is similar to many publicly available prediction systems in its overall goal of predicting MHC binding peptides, there are significant differences in both functionality and methodology. MULTIPRED1 predicts promiscuous peptides binding to multiple HLA proteins with one model per HLA supertype. Since HLA

proteins are highly polymorphic, promiscuous peptides that bind more than one HLA protein are prime targets for vaccine and immunotherapy development because they are relevant to higher proportions of the human population. MULTIPRED1 provides several functions which are not available in other prediction systems, such as model building by users, accuracy evaluation function and consensus prediction function (details are available in Appendix C). The pathway from epitopes to vaccine development is lengthy and cost-intensive, involving exhaustive experiments. Prediction tools like MULTIPRED1 can help to minimize the number of experiments required for mapping of promiscuous T-cell epitopes.

Chapter 6 Hotspot Hunter: a computational system for large-scale screening and selection of candidate immunological hotspots in pathogen proteomes

An ideal epitope-based vaccine needs to contain a minimal number of epitopes that cover a vast majority of human population [195, 220, 221]. Promiscuous peptides that bind more than one HLA allele are prime targets for vaccine and immunotherapy development because they are relevant to higher proportions of the human population. The presence of immunological hotspots, regions of high density of promiscuous peptides, has been observed in several antigens, such as SARS coronavirus nucleocapsid [177], HIV-1 proteins [222-224], and Chlamydia trachomatis outer membrane protein [225]. These clusters may be exploited to facilitate the development of epitope-based vaccines by selecting a small number of hotspots that can elicit all of the required T-cell functions [177].

6.1. Introduction

Nearly all publicly available promiscuous T-cell epitope prediction servers perform prediction for a single protein sequence per submission. Given the large size of pathogen proteomes, it is painstaking to integrate the individual prediction results, and therefore impractical to use these tools for large-scale systematic study of promiscuous epitopes, which is necessary for a global view of the localizations of the epitopes in the proteome and analysis of their antigenic diversity. Large-scale screening for HLA supertype-restricted epitopes and antigenic diversity analysis are important for dealing

with the diversity of HLA alleles in the human population and viral antigens [226, 227].

Hotspot Hunter, a web-based computational system for large-scale screening and selection of candidate immunological hotspots in pathogen proteomes through analysis of antigenic diversity has been developed. It predicts hotspots specific to four common HLA supertypes, namely HLA class I A2, A3, and B7, and class II DR.

It has been reported that combining predictions by several methods results in greater accuracy [228-230]. As it was shown in section 5.3.10, ANN and SVM models perform better than HMM models in most cases. Thus, Hotspot Hunter uses ANN and SVM methods as predictive engines. Hotspot Hunter takes advantage of the “collective wisdom” of the both methods by integrating their prediction results. Then the results are presented to users in a succinct and easily understood format. Users can predict hotspots from a single protein sequence, or from a set of aligned protein sequences representing the pathogen proteome. The latter feature provides a global view of the localizations of the hotspots in the proteome set, enabling analysis of antigenic diversity and shift of hotspots across protein variants. Antigenic variation, particularly in pathogens, may result in disappearance of some hotspots and appearance of new hotspots across variants [227, 231]. The system provides a target selection feature for multiple sequence queries to identify candidate peptides for the formulation of an epitope-based vaccine that could be effective against multiple variants of the pathogen and applicable to a large proportion of the human population [221]. Hotspot Hunter is publicly accessible at <http://antigen.i2r.a-star.edu.sg/hh/>.

6.2. Material and Methods

6.2.1 Training datasets

The predominant length of peptides that bind HLA class I molecules (HLA-A, -B, and -C) is nine amino acids [174]. HLA class II molecules (HLA-DR) bind longer peptides although through a nine amino acid long binding core [232]. The training datasets for HLA-A2, A3 and B7 supertypes are Dataset1, Dataset2, and Dataset3 described in section 5.2.1. The HLA-DR training dataset had 2396 (448 binders and 1948 non-binders) 9-mer peptides related to six HLA-DR alleles (HLA-DRB1*0101, *0301, *0701, *0801, *1101 and *1501). They were from experiments conducted on 340 15-mers to measure their binding affinities to each of the six HLA-DR alleles. The 15-mers came from three protein sources: MEL40 (aka SSX2, NCBI Accession: AAH16957.1, human melanoma cancer), SSP2 (malaria sporozoite surface protein 2, NCBI Accession: Q01443, *Plasmodium yoelli*), HCV (Hepatitis C virus genome polyprotein, NCBI Accession: P26663). Since the data originated from a single set of experiments, with relatively consistent experimental conditions, it is reasonable to expect that the DR training dataset is of higher quality relative to datasets aggregated from multiple sources.

6.2.2 Prediction engine

Three-layer BP networks with various architectures have been trained using the HLA-DR dataset. We found that they either took a long time to converge or do not converge at all. It was shown in [233] that 4-layered feedforward neural networks have greater classification capabilities than 3-layered feedforward neural networks. Thus, 4-layer BP

networks with various architectures have been trained on the HLA-DR data. The ANN that had the best performance was a 4-layer BP network (268-2-4-1) with a hyperbolic tangent sigmoid activation function as shown in equation (2.16) between the two hidden layers and a sigmoid activation as shown in equation (2.15) function between the second hidden layer and the output. The value of momentum was 0.9. The learning rate was not fixed but changed according to the learning process to achieve faster convergence. The initial learning rate was 0.004. In each training epoch, the sum of square errors was compared with that of the previous epoch. If the sum of square errors increased to more than 1.005 times the previous one, the learning rate was decreased to 0.7 times that of the initial learning rate. If the sum of square errors decreased to less than 0.98 times the previous one, the learning rate was increased to 1.05 times that of the initial learning rate.

For HLA-DR SVM models, the selection of kernel functions and parameter values was done through sampling and comparison of performance. The SVM model that had the best performance employed a 2-degree polynomial kernel function for prediction of peptide binding to the HLA-DR supertype.

A)

B)

Figure 6.1 Processing steps in prediction of T-cell epitope hotspots. A) For HLA class I supertypes. B) For HLA class II supertypes.

Flowcharts describing the steps involved in identifying hotspots in a single input protein are shown in Figure 6.1. Each input protein sequence to Hotspot Hunter is truncated into overlapping 9-mer peptide sequences with an eight amino acid overlap and analysis is carried out on each individual 9-mer. To predict immunological hotspots (regions of high concentration of 9-mer promiscuous binders), we have developed two scoring schemes for HLA class I and class II supertypes. The scheme for HLA class I supertypes is based on high scoring individual 9-mers within a window of 30 amino acids [199, 234], and the scheme for HLA class II supertype is based on average scores of individual 9-mers within a window of 15 amino acids. The selection of window lengths was based on a trial-and-error based heuristics. Window lengths of 15, 20, 25 and 30 were explored and the results were compared with the representative experimental results. The window length of 30 amino acids was found to be best for class I predictions and the window length of 15 amino acids was best for class II predictions. A region in a protein sequence is considered as an immunological hotspot specific to a HLA supertype only if its predicted binding strength is above the threshold defined for the supertype. The threshold for each supertype is the sum of the thresholds of corresponding ANN and SVM models, which are selected based on experimental binding data [199, 212].

6.3. Model Validation

Ten-fold cross-validation results showed that the AROC values for A2, A3, B7, and DR ANN models were 0.83, 0.83, 0.84, and 0.85, respectively. AROC values of A2, A3, B7, and DR SVM models in 10-fold cross-validation were 0.91, 0.95, 0.90, and 0.80, respectively.

6.3.1 Validation of HLA-A2 and A3 prediction models using experimental results

The prediction performance of models for HLA-A2 and -A3 supertypes was also validated using experimental results from a systematic study of HPV 16 E6 (NCBI Accession: P03126) and E7 (NCBI Accession: P03129) proteins. Experimentally identified HLA-A*0201 and HLA-A*0301 binding peptides in HPV E6 and E7 are listed in Table 5.27 and Table 5.33. In protein E6, E6-7, E6-18 and E6-26 cluster together and form a HLA-A2 restricted hotspot region, E6 7-34. In protein E7, HLA-A2 restricted hotspots formed by the seven HLA-A2 binders are E7 7-20 (E7-7, E7-11 and E7-12) and E7 82-94 (E7-82, E7-85 and E7-86). In protein E6, HLA-A3 restricted hotspots formed by the nine HLA-A*0301 binders are E6 33-67 (E6-33, E6-42 and E6-59), E6 75-101 (E6-75, E6-89 and E6-93), and E6 125-151 (E6-125 and E6-143). There is no HLA-A3 restricted hotspot in protein E7 as there is only one HLA-A*0301 binder in HPV E7.

As described in Appendix D, graphs providing an overall view of the binding capacities of input sequences can be generated by Hotspot Hunter using “Plot Binding Strength” function on the result pages of “Single Sequence Query”. The X-axis represents the start position of a peptide (window size of 15-amino acids for DR supertype and 30-amino acids for A2, A3 and B7 supertypes) and Y-axis represents the predicted score (ANNSUMavg+SVMSUMavg for DR supertype and ANNScore4+SVMScore4 for A2, A3 and B7 supertypes) of the peptide.

Figure 6.2A shows a plot of the binding scores for HPV type 16 E6 protein with each dot (in blue or red color) representing a 30-mer peptide. For example, the first 30-amino acid peptide is plotted at position 1 with a score of 79.17. Its coordinate is $(x, y) = (1, 79.17)$. The red dots represent the predicted 30-mer binders and the red bars represent the

predicted hotspots. The red bars are longer than the regions formed by the red dots because the red dots represent the start positions of 30-mers while the red bars span the full length of the predicted hotspots. The pink bars at the bottom of the graphs represent the locations of the experimentally identified hotspots. The predicted HLA-A2 restricted immunological hotspot in HPV type 16 E6 protein is 8-47 as shown in Figure 6.2A, which covers 96% of the experimental identified HLA-A2 restricted hotspot, E6 7-34.

The predicted HLA-A3 restricted immunological hotspots in HPV type 16 E6 protein are 27-62, 72-118, and 125-158, as shown in Figure 6.2B. The first predicted hotspot partially (86%) covers the experimental identified HLA-A3 restricted hotspot, E6 33-67. The latter two predicted hotspots cover two experimental identified HLA-A3 restricted hotspots, E6 75-101 and E6 125-151.

As shown in Figure 6.3, the predicted HLA-A2 restricted immunological hotspot in HPV type 16 E7 is 61-98, which fully covers the experimental identified HLA-A2 restricted T-cell epitope hotspot, E7 82-94. The other experimental identified HLA-A2 restricted T-cell epitope region, E7 7-20, was missed out by the predictions.

There is no HLA-A3 restricted immunological hotspot predicted in HPV type 16 E7 protein by Hotspot Hunter, which corresponds to the experimental results.

6.3.2 Validation of HLA-DR prediction model using experimental results

The prediction performance for HLA-DR hotspots was validated using experimental results from systematic binding studies of overlapping peptides from myelin oligodendrocyte glycoprotein (MOG) (NCBI Accession: CAA88109) [235] and hepatitis C virus 1B protein (HCV 1B) (NCBI Accession: AAB00216) [106]. Multiple sclerosis

(MS) is an inflammatory and demyelinating disease of the central nervous system. MOG induces a type of experimental autoimmune encephalomyelitis in animals which is very MS-like. Thus, T-cell reactivity pattern analysis of MOG is relevant to MS pathogenesis. Experimentally identified HLA-DR restricted peptides in the mature sequence of MOG form two hotspots, MOG 73-96 and MOG 121-216. There are three experimentally identified T-cell epitope regions in HCV 1B protein, HCV 1B 29-52, HCV 1B 95-108 and HCV 1B 118-173.

The prediction results for MOG were reported in Appendix D. As shown in Figure D.1 B, four HLA-DR restricted hotspots, 38-56, 79-93, 121-163 and 174-211, were predicted in MOG. Three of the four predicted hotspots, 79-93, 121-163 and 174-211, corresponded to the two experimentally identified T-cell epitope regions, 73-96 and 121-216 as shown in Figure D.1 C. Prediction 38-56 was a false positive.

As shown in Figure 6.4, three HLA-DR restricted hotspots, 30-47, 130-147 and 172-191, were predicted in HCV 1B. The predicted hotspot, 30-47, partially (75%) covers the experimentally identified T-cell epitope regions, 29-52. The predicted hotspot, 130-147, partially covers (32%) the experimentally identified T-cell epitope regions, 118-173. The last prediction, 172-191, is a false positive and the experimentally identified T-cell epitope regions, 95-108, is missed out by the predictions.

Additional validation study using secretory aspartyl proteinase 2 (Sap2), a major protein known to induce immune response during *Candida* infection in human, showed that Hotspot Hunter accurately identified two of the three HLA-DR restricted immunological regions experimentally identified using PBMC proliferation and IL-2 ELISpot assays [236].

Figure 6.2 A) Predicted HLA-A2 restricted immunological hotspots in HPV type 16 E6 protein. B) Predicted HLA-A3 restricted immunological hotspots in HPV type 16 E6 protein. The red dots represent the start positions of the predicted 30-mer binders and the red bars represent the predicted hotspots. The pink bars represent the positions of experimentally identified hotspots.

Figure 6.3 Predicted HLA-A2 restricted immunological hotspots in HPV type 16 E7 protein. The red dots represent the start positions of the predicted 30-mer binders and the red bars represent the predicted hotspots. The pink bars represent the positions of experimentally identified hotspots.

Figure 6.4 Predicted HLA-DR restricted immunological hotspots in HCV 1B protein. The red dots represent the start positions of the predicted 15-mer binders and the red bars represent the predicted hotspots. The pink bars represent the positions of experimentally identified hotspots.

6.3.3 Independent validation of HLA-A2 prediction model

Stoloff and Caparros-Wanderley recently proposed a new candidate synthetic influenza vaccine containing multiple conserved T-cell epitopes (polyepitope vaccine) naturally presented by human cells infected with non-related influenza strains [237]. They selected six conserved regions from influenza proteome ranging from 21 to 35 amino acid in length, each containing at least five HLA-A*0201 T-cell epitopes (Table 6.1). These sequences are conserved in at least 70% of the reported influenza isolate sequences. Immunization of humanized transgenic mice with a vaccine formulation containing these peptides induced specific HLA-A*0201 mediated CD8+ T-cell responses and offered a level of protection of mice from influenza challenge. The challenge of mice with a deadly dose of influenza resulted in 100% death rate, while immunization reduced the death rate to 40%.

We selected representative influenza sequences containing the six peptides (Swiss-Prot accessions: P35937 – M1A, Q71UK7 – M1B, P26083 – NPA, P04666 – NPB, P16509 – PB1, and P35938 – M2) used in the Stoloff and Caparros-Wanderley vaccine and analyzed them by the Hotspot Hunter. Five of the predicted hotspots contain full-length Stoloff and Caparros-Wanderley peptides (highlighted in red in Table 6.2). In the protein Q71UK7 – M1B, the predicted hotspot 122-151 contains 28 of the 35 amino acids from the M1B peptide which is at position 124-158 (80% overlap) as shown in Table 6.2 B. The validation results show that Hotspot Hunter is able to make highly accurate predictions of functionally important immunogenic regions.

155

Table 6.1 The six conserved Influenza protein fragments containing multiple T-cell epitopes selected by Stoloff and Caparros-Wanderley.

Name	Consensus Sequence	Position	Length
M1A	DLEALMEWLKTRPILSPLTKGILGFVFTLTVP	36-67	32
M1B	LLYCLMVMYLNPGNYSMQVKLGTLCALCEKQASHS	124-158	35
NPA	DLIFLARSALILRGSVAHKSC	255-275	21
NPB	PGIADIEDLTLLARSMVVVRP	306-326	21
PB1	LLIDGTASLSPGMMMGMFNMNSTVLGVLSILNLGQ	395-428	34
M2	IIGILHLILWILDRLFFKCIYRLF	32-55	24

Table 6.2 HLA-A2 restricted hotspots predicted by Hotspot Hunter in the six Influenza proteins. The six regions shown in Table 6.1 are highlighted in red. A) P35937 – M1A; B) Q71UK7 – M1B; C) P26083 - NPA; D) P04666 - NPB; E) P16509 - PB1; F) P35938 – M2.

A)

Position	Sequence	Binding Strength	Length
34-80	GKNTDLEALMEWLKTRPILSPLTKGILGFVFTLTVP SERGLQRRRFV	92.50	47
102-131	KLKREITFHGAKEIALSYSAGALASCMGLI	88.00	30

B)

Position	Sequence	Binding Strength	Length
37-80	FDLDSALEWIKNKRCLTDIQKALIGASICFLKPK DQERKRRFIT	90.41	44
186-222	AKTMNGMGKGEDVQKLAEELQSNIGVLRSLG ASQKNG	88.43	37
122-151	SALLLYCLMVMYLNPGNYSMQVKLGTLCALC	88.42	30

C)

Position	Sequence	Binding Strength	Length
237-326	AMMDQVRESRNPNGNAEIEDLIFLARSALILRGSV AHKSCLPACVYGLAVASGHDFEREGLYSLVGIEPF	90.11	90
39-68	KLLQNSQVFSLRPNENPAHKS FYIQMCTELKLSDYEGRLIQNSITIERMVL	88.40	30

D)

Position	Sequence	Binding Strength	Length
294-349	KALVDQVIGSRNPGIADIEDLTLLARSMVVVRPSV ASKVVLPISIYAKIPQLGFNV	93.44	56
412-442	SALKCKGFHVPAPKEQVEGMGAALMSIKLQFW	88.95	31

156

E)

Position	Sequence	Binding Strength	Length
386-442	RKRIEKIRP LLIDGTASLSPGMMMGMFNMLSTVLG VSILNLGQ KRYTRTTYWWDGLQ	94.18	57

F)

Position	Sequence	Binding Strength	Length
10-66	PIRNEWGCRCNDSSDPLVVAAS IIGILHLILWILDRLE FKCIYRLF KHGLKRGPE	97.91	57

6.4. Discussion

Experimental approaches for identification of T-cell epitopes are laborious and costly, and thus are not applicable for large-scale screening across multiple HLA alleles and pathogen proteomes. Several publicly available online computational systems have been developed for the prediction of peptides binding to HLA alleles and supertypes, such as SYFPEITHI [238] based on binding motif, BIMAS [97], ProPred1 [239], and PEPVAC [135] based on quantitative matrices, SMM [103] based on stabilized matrix method, MHCpred [171] based on a multivariate statistical method, SVMHC [115] based on SVM, and NetCTL [134] based on an integrative approach combining MHC class I binding, TAP transport efficiency, and proteasomal cleavage.

Hotspot Hunter is different from the above mentioned online servers in several aspects. First, the prediction engine combines the strengths of the ANN and SVM methods for better and more robust prediction performance; second, it presents users with predicted immunological hotspots, which is a region of high concentration of predicted HLA binding peptides; third, it is suitable for systematic studies of a large set of pathogen

proteomes as it can concurrently analyze multiple sequences and present a map providing a global view of their localizations in the proteome; finally, it provides a utility for selecting candidate experimental targets based on antigenic diversity analysis. Therefore, Hotspot Hunter is a new generation computational tool aiding in epitope-based vaccine design. A customized version of Hotspot Hunter has been integrated into some of our in-house specialized databases, such as the Tumor Antigen Database (research.i2r.a-star.edu.sg/Templar/DB/cancer_antigen/) and CandiVF - Candida albicans Virulence Factor Database [240], for prediction of T-cell epitope hotspots.

Chapter 7 Contribution and Future Work

Development of epitope-based vaccines is progressing rapidly. Induction of epitope-specific T-cell responses can help in fighting diseases for which no conventional vaccines exist [241]. It has been reported that vaccines based on immune epitopes fighting both viruses and cancers induce protection in animal models as well as humans [231, 242-244]. Selection of peptides for candidate vaccines is one of the greatest obstacles in the development of epitope-based vaccines that have broad population coverage for excessive polymorphism of HLA molecules.

Computer models are used to complement laboratory experiments and thereby speed up vaccine targets discovery. The number of laboratory experiments for T-cell epitope mapping can be minimized by the judicious use of experiments aimed at developing and validating computer models. These models can then be used to perform large-scale computer simulations rapidly and inexpensively. The hypotheses generated from these simulations can then be retested in the laboratory to confirm their applicability to real-life immunology. In this work, we have developed a suite of computational tools aiding fast identification of T-cell epitopes with board population coverage.

7.1. Contribution

The BALB/c mouse is an important laboratory model organism and widely used in immunological studies and vaccine research. We developed $\text{PRED}^{\text{BALB/c}}$, a computational system that predicts peptide binding to the H2^d alleles of the BALB/c mouse. $\text{PRED}^{\text{BALB/c}}$

marks a new direction in predictive modeling of MHC binding peptides and T-cell epitopes. The main advantage is that PRED^{BALB/c} focuses on a complete organism and its predictions represent a complete set of predicted targets of T-cell immune responses. The focus on the complete set of MHC alleles is closer to studies involving laboratory animals. This approach provides a more complete view of the immune responses of an organism. PRED^{BALB/c} is, therefore, useful for the analysis of immunization regimens and the study of immunological diseases.

Combined computational methods that integrate multiple critical steps of the MHC class I antigen processing pathway have been proposed as a supporting methodology for prediction of T-cell epitopes and several applications modeling the pathway have been reported. We developed PRED^{TAP}, a computational system that predicts peptides binding to hTAP. PRED^{TAP} can be integrated with systems for predictions of HLA class I binding peptides to achieve greater accuracy in T-cell epitope prediction. However, these combined predictions need to be taken with a dose of caution. Alternative pathways for both proteolytic degradation [192] and TAP transport [193] have been reported. In some cases TAP-deficient individuals have normal immune responses [194], suggesting that TAP-independent immune responses are sufficient to provide effective protection from some intracellular pathogens. Nevertheless, the proteasome-TAP-MHC class I pathway is responsible for 90-97% of expression of peptide/MHC Class I complexes and therefore is critical for the identification of target epitopes for immunotherapies and vaccines.

Many bioinformatics tools have been developed for prediction of HLA binding peptides. Most of them predict peptide binding to a single HLA molecule instead of promiscuous HLA binders. Using a novel data representation that combines sequences of

HLA molecules and peptides to capture the interaction between them, together with ANN, HMM and SVM methods, we developed a computation system, MULTIPRED1. It focuses on predicting promiscuous HLA binders relevant to a large proportion of human population. The system demonstrated reliable performance for prediction of 9-mer HLA-A2, A3, B7 and DR binders. The main difference between our prediction models and the models employed in other computational tools for prediction of HLA binders is that we built a single model for prediction of peptides that bind HLA alleles from an entire supertype. In addition, our models can predict peptide binding to HLA-A2, A3, B7 and DR alleles for which no binding data are available. For example, as shown in Table 5.39, the HLA-A2 SVM model performed reasonably well in predicting A*6901 binding peptides (AROC=0.81), when validated using the Peters dataset [178]. This is in spite of the fact that only four peptides are related to A*6901 in Dataset1 (Table 5.1), the dataset that was used for training of the model. However, when the number of peptides related to a HLA molecule in a training dataset is small or the peptides are biased and do not represent the diverse pattern of binding peptides to the molecule, the resulting prediction models may perform poorly, as is the case for the HLA-A3 SVM model when predicting binding peptides for A*6801 (Table 5.40). This emphasizes the need for use of representative datasets for development of MHC-binding prediction servers. While the accuracy of predictions will improve with the addition of new data to the training sets, it is important to note that about a quarter of MHC class I binding peptides lack canonical binding motifs. These peptides are underrepresented in the current datasets, which are used both for training and testing, and therefore the values of accuracy of predictive methods are likely to be somewhat lower than reported, across all comparative studies.

Most of publicly available T-cell epitope prediction servers perform prediction for a single protein sequence per submission. Given the large size of pathogen proteomes, it is painstaking to integrate the individual prediction results, and therefore impractical to use these tools for large-scale systematic study of promiscuous epitopes, which is necessary for a global view of the localizations of the epitopes in the proteome and analysis of their antigenic diversity. In view of the importance of dealing with the diversity of HLA alleles in the human population and viral antigens [226, 227], we developed Hotspot Hunter, a web-based computational system for large-scale screening and selection of candidate immunological hotspots in pathogen proteomes through analysis of antigenic diversity. It allows screening and selection of hotspots specific to four common HLA supertypes, namely HLA class I A2, A3, B7, and class II DR.

It has been reported that combining predictions by several methods results in greater accuracy [228-230]. Thus, a soft computing approach was applied to integrate the prediction results produced by both ANN and SVM methods in Hotspot Hunter. Hotspot Hunter is different from other online servers in several aspects. First, its prediction engine combines the strengths of the ANN and SVM methods for better and more robust prediction performance; second, it presents users with predicted immunological hotspots, which are regions of high concentration of predicted promiscuous HLA binding peptides; third, it is suitable for systematic studies of a large set of pathogen proteomes as it can concurrently analyze multiple sequences and present a map providing a global view of their localizations in the proteome; finally, it provides a utility for selecting candidate experimental targets based on antigenic diversity analysis. Because it employs a new concept of T-cell epitope hotspots and enables their identification within multiple variants

of protein antigens, the Hotspot Hunter is a new generation computational tool aiding epitope-based vaccine design.

The research work presented here has been published or accepted for publication as journal articles, conferences papers, and book chapters. They include the development of PRED^{BALB/c} (Chapter 3) [245], PRED^{TAP} (Chapter 4) [246], MULTIPRED1 (Chapter 5) [199], and Hotspot Hunter (Chapter 6) [247]. A review article focusing on computational methods for *in Silico* T-cell epitope prediction has been submitted to an immunological journal.

7.2. Future work

Analysis focusing on a complete organism provides a more complete view of the immune responses of an organism. The focus on the complete set of MHC alleles is closer to studies involving laboratory animals. This approach reported in Chapter 3 points a new direction in predictive modeling of MHC-binding peptides and T-cell epitopes. Further developments would include modeling of other laboratory animal organisms.

Computational systems for prediction of vaccine targets in human as well as other organisms have been developed. It was found one method may outperform others on some molecules, but not one method always performs best in all cases. It has been reported that combining predictions by several methods results in greater accuracy [228-230]. Thus, soft computing approaches combining multiple prediction methods should be explored for more accurate T-cell epitope prediction. A preliminary method of integrating ANN and SVM methods has been reported in Chapter 6. Future work would include development of more sophisticated approaches to integrate multiple prediction

methods.

The methods reported in this thesis belong to sequence-based methods. Recently hybrid methods based on structure and sequence methods were reported for MHC binding peptide prediction. One is named adaptive double threading, an improved protein threading approach which uses not only MHC/peptide complex structure data but also additional information from MHC sequences and binding affinities between MHC and peptides [129]. The model built significantly outperforms the standard threading approach in binding energy prediction because of the additional information used. Another method, termed DynaPred, is a SVM-trained, quantitative matrix-based method with the features of the scoring matrix being energy terms retrieved from MD simulations. [145]. Further developments would include combining information both from sequence and structure for more accurate prediction of T-cell epitopes.

Large-scale screening for HLA supertype-restricted epitopes and antigenic diversity analysis are important for dealing with the diversity of HLA alleles in the human population and viral antigens. The presence of immunological hotspots, regions of high density of promiscuous peptides, has been observed in several antigens, such as SARS coronavirus nucleocapsid [177], HIV-1 proteins [222-224], and *Chlamydia trachomatis* outer membrane protein [225]. The computational system reported in Chapter 6 paves the way for future developments in large-scale T-cell epitope screening in large sets of pathogen proteomes, pathogen antigenic diversity analysis, and identification of immunological hotspots.

Author's Publication

1. **Zhang,G.L.**, Khan,A.M., Srinivasan,K.N., Heiny,A.T., Lee,K.X., Kwoh,C.K., August,J.T., Brusic,V. (2008) Hotspot Hunter: a computational system for large-scale screening and selection of candidate immunological hotspots in pathogen proteomes. *BMC Bioinformatics* **9**(Suppl 1), S19.
2. **Zhang,G.L.**, Bozic, I., Kwoh,C.K., August,J.T. and Brusic,V. (2007) Prediction of Supertype-Specific HLA Class I Binding Peptides Using Support Vector Machines. *J Immunol Methods* **320**, 143-154.
3. **Zhang,G.L.**, Petrovsky,N., Kwoh,C.K., August,J.T. and Brusic,V. (2006) PRED^{TAP}: a system for prediction of peptide binding to the human transporter associated with antigen processing. *Immunome Res.* **2**(1), 3.
4. **Zhang, G.L.**, Khan, A.M., Srinivasan, K.N., August, J.T. and Brusic, V. (2005) Neural Models for predicting viral vaccine targets. *J Bioinform Comput Biol.* **3**(5), 1207-1225.
5. **Zhang,G.L.**, Khan,A.M., Srinivasan,K.N., August,J.T. and Brusic,V. (2005) MULTIPRED: a computational system for prediction of promiscuous HLA binding peptides. *Nucleic Acids Res.* **33**, W172 - W179.
6. **Zhang,G.L.**, Srinivasan,K.N., Veeramani,A., August,J.T. and Brusic,V. (2005) PRED^{BALB/c}: a system for prediction of peptide binding to the H2^d molecules, a haplotype of the BALB/c mouse. *Nucleic Acids Res.* **33**, W180 - W183.
7. Srinivasan,K.N., **Zhang,G.L.**, Khan,A.M., August,J.T. and Brusic,V. (2004) Predictions of Class I T-cell epitopes: Evidence of presence of immunological hot

spots inside antigens. *Bioinformatics*. **20 Suppl 1**, i297-i302.

8. Handoko,S.D., Kwoh,C.K., Ong,Y.S., **Zhang,G.L.** and Brusic,V. (2006) Extreme Learning Machine for Predicting HLA-Peptide Binding. Lecture Notes in Computer Science, Springer, 3973, 716-721.
9. Bozic,I., **Zhang,G.L.** and Brusic,V. (2005) Predictive Vaccinology: Optimisation of Predictions Using Support Vector Machine Classifiers. Lecture Notes in Computer Science, Springer, 3578, 375-381.
10. **Zhang,G.L.**, Tong,J.C., Zhang,Z.H., Zheng,Y., August,J.T., Kwoh,C.K. and Brusic,V. (2006) Computational models for identifying promiscuous HLA-B7 binders based on information theory and support vector machine. ICBPE, Singapore.
11. **Zhang,G.L.**, Kwoh,C.K., August,J.T. and Brusic,V. (2006) Performance Evaluation of MULTIPRED1 on Prediction of MHC Class I Binders. ICBPE, Singapore.

Bibliography

1. Fabbri, M., C. Smart, and R. Pardi, *T lymphocytes*. Int J Biochem Cell Biol, 2003. **35**: p. 1004-1008.
2. Bjorkman, P.J., et al., *The foreign antigen binding site and T cell recognition regions of class I histocompatibility antigens*. Nature, 1987. **329**: p. 512-518.
3. Watts, C. and S. Amigorena, *Phagocytosis and antigen presentation*. Semin Immunol, 2001. **13**: p. 373-379.
4. Hudson, A.W. and H.L. Ploegh, *The cell biology of antigen presentation*. Exp Cell Res, 2002. **272**: p. 1-7.
5. Strehl, B., et al., *Interferon-gamma, the functional plasticity of the ubiquitin-proteasome system, and MHC class I antigen processing*. Immunol Rev, 2005. **207**: p. 19-30.
6. Donnes, P. and O. Kohlbacher, *Integrated modeling of the major events in the MHC class I antigen processing pathway*. Protein Sci, 2005. **14**(8): p. 2132-40.
7. Gromme, M. and J. Neefjes, *Antigen degradation or presentation by MHC class I molecules via classical and non-classical pathways*. Mol Immunol, 2002. **39**(3-4): p. 181-202.
8. Petrovsky, N. and V. Brusica, *Virtual models of the HLA class I antigen processing pathway*. Methods, 2004. **34**(4): p. 429-35.
9. Williams, T.M., *Human leukocyte antigen gene polymorphism and the histocompatibility laboratory*. J Mol Diagn, 2001. **3**: p. 98-104.
10. Sette, A. and J. Sidney, *HLA supertypes and supermotifs: a functional perspective on HLA polymorphism*. Curr Opin Immunol, 1998. **10**(4): p. 478-482.
11. Sette, A., et al., *Optimizing vaccine design for cellular processing, MHC binding and TCR recognition*. Tissue Antigens, 2002. **59**(6): p. 443-51.
12. Sette, A. and J. Sidney, *Nine major HLA class I supertypes account for the vast preponderance of HLA-A and -B polymorphism*. Immunogenetics, 1999. **50**(3-4): p. 201-212.
13. Lund, O., et al., *Definition of supertypes for HLA molecules using clustering of specificity matrices*. Immunogenetics, 2004. **55**(12): p. 797-810.
14. Hertz, T. and C. Yanover, *Identifying HLA supertypes by learning distance functions*. Bioinformatics, 2007. **23**(2): p. e148-55.
15. Sidney, J., et al., *Practical, biochemical and evolutionary implications of the discovery of HLA class I supermotifs*. Immunol Today, 1996. **17**: p. 261-266.
16. Gulukota, K., *Skipping a step with neural nets*. Nat Biotechnol., 1998. **16**(8): p. 722-723.
17. Brusica, V. and J. Zeleznikow, *Computational binding assays of antigenic peptides*. Letters in Peptide Science, 1999. **6**(5): p. 313-324.
18. Robinson, J., et al., *The IMGT/HLA and IPD databases*. Hum Mutat, 2006. **27**(12): p. 1192-9.
19. Robinson, J., et al., *IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex*. Nucleic Acids Res., 2003. **31**(1): p. 311-314.
20. Brusica, V. and N. Petrovsky, *Immunoinformatics and its relevance to*

- understanding human immune disease*. Expert Rev. Clin. Immunol., 2005. **1**: p. 145-157.
21. Brusica, V., G. Rudy, and L.C. Harrison, *MHCPEP: a database of MHC-binding peptides*. Nucleic Acids Res., 1994. **22**: p. 3663-3665.
22. Rammensee, H.G., et al., *SYFPEITHI: database for MHC ligands and peptide motifs*. Immunogenetics, 1999. **50**(3-4): p. 213-219.
23. Reche, P.A., et al., *EPIMHC: a curated database of MHC-binding peptides for customized computational vaccinology*. Bioinformatics, 2005. **21**(9): p. 2140-2141.
24. Blythe, M.J., I.A. Doytchival, and D.R. Flower, *JenPep: a database of quantitative functional peptide data for immunology*. Bioinformatics, 2002. **18**(3): p. 434-439.
25. Peters, B., et al., *The immune epitope database and analysis resource: from vision to blueprint*. PLoS Biol, 2005. **3**(3): p. e91.
26. Yu, K., et al., *Methods for prediction of peptide binding to MHC molecules: a comparative study*. Mol Med, 2002. **8**(3): p. 137-48.
27. Zadeh, L.A., *Applied Soft Computing - Foreword*. Applied Soft Computing, 2001. **1**: p. 1-2.
28. Zadeh, L.A., *Fuzzy Logic and Soft computing: Issues, Contention and Perspectives*, in *IIZUKA'94: 3rd International Conference on Fuzzy Logic, Neural Nets and Soft Computing*. 1994: Iisuka, Japan. p. 1-2.
29. Dote, Y. and S.J. Ovaska, *Industrial Applications of Soft Computing: A Review*. PROCEEDINGS OF THE IEEE, 2001. **89**(9): p. 1243-1265.
30. Zadeh, L.A., *Fuzzy Sets*. Information and Control, 1965. **8**(3): p. 338-353.
31. Holland, J.H., *Adaptation in Natural and Artificial Systems*. 1975, Cambridge, MA: MIT press.
32. Zurada, J.M., *Introduction to Artificial Neural Systems*. 1999, St. Paul, MN, USA: PWS Publishing Co.
33. Eddy, S.R., *Profile hidden Markov models*. Bioinformatics, 1998. **14**(9): p. 755-763.
34. McInerney, M. and A.P. Dhawan. *Use of genetic algorithms with backpropagation in training of feedforward neural networks*. in *Proc. IEEE Int. Conf. on Neural Networks, IEEE-ICNN'93*. 1993. San Francisco, CA.
35. Bonissone, P.P., *Soft computing: the convergence of emerging reasoning technologies*. Soft Comput, 1997. **1**: p. 6-18.
36. Kitano, H. *Empirical studies on the speed of convergence of neural networks training using genetic algorithms*. in *Eighth National Conference in Artificial Intelligence (AAAI'90)*. 1990. Boston, MA.
37. Arabshahi, P., et al. *Fuzzy Control of Backpropagation*. in *First IEEE International Conference on Fuzzy Systems*. 1992. San Diego, CA.
38. Herrera, F. and M. Lozano, *Adaptive Genetic Algorithms Based on Fuzzy Techniques*, in *IPMU'96*. 1996: Granada, Spain. p. 775-80.
39. Karr, C.L., *Fuzzy control of ph using genetic algorithms*. IEEE Transactions on Fuzzy Systems, 1993. **1**: p. 46-53.
40. Takagi, H., *Fusion Technology of Fuzzy Theory and Neural Networks - Survey and Future Directions*. Proceedings of the 1990 International Conference on Fuzzy

- Logic and Neural Networks, 1990: p. 13-26.
41. Takagi, H., *Fusion technology of neural networks and fuzzy systems: a chronicled progression from the laboratory to our daily lives*. International Journal of Applied Mathematics and Computer Science, 2000. **10**(4): p. 647–673.
42. Renders, Y.M. and H. Bersini, *Hybridizing Genetic Algorithms with Hill Climbing Methods for Global Optimization: Two Possible Ways*. Proceedings of the first IEEE Conference on Evolutionary Computation, 1994. **1**: p. 312-17.
43. Won, K.-J., A. Prugel-Bennett, and A. Krogh, *Training HMM structure with genetic algorithm for biological sequence analysis*. Bioinformatics, 2004. **20**(18): p. 3613-19.
44. Rajapakse, M., et al., *Deriving matrix of peptide interactions in diabetic mouse by genetic algorithm*. Lect Notes Comp Sci, 2005. **3578**: p. 440–447.
45. Kruse, R. and A. Klose, *Recent advances in exploratory data analysis with neuro-fuzzy methods*. 2004. **8**(6): p. 381-382.
46. Paetz, J., *Descriptor vector redesign by neuro-fuzzy analysis*. Soft Comput, 2005. **9**.
47. Zhao, Y. and C.K. Kwoh, *Fast leave-one-out evaluation for dynamic gene selection*. Soft Comput, 2005. **9**.
48. Ada, G., *Progress towards achieving new vaccine and vaccination goals*. Intern Med J, 2003. **33**(7): p. 297-304.
49. Durrant, L.G. and J.M. Ramage, *Development of cancer vaccines to activate cytotoxic T lymphocytes*. Expert Opin Biol Ther, 2005. **5**(4): p. 555-563.
50. Verhagen, J., et al., *Targets in allergy-directed immunotherapy*. Expert Opin Ther Targets, 2005. **9**(2): p. 217-224.
51. Quintana, F.J. and I.R. Cohen, *DNA vaccines coding for heat-shock proteins (HSPs): tools for the activation of HSP-specific regulatory T cells*. Expert Opin Biol Ther, 2005. **5**(4): p. 545-554.
52. Wisniewski, T. and B. Frangione, *Immunological and anti-chaperone therapeutic approaches for Alzheimer disease*. Brain Pathol, 2005. **15**(1): p. 72-77.
53. Hilleman, M.R., *Vaccines in historic evolution and perspective: a narrative of vaccine discoveries*. Vaccine, 2000. **18**: p. 1436-1447.
54. Brusica, V. and D.R. Flower, *Bioinformatics tools for identifying T-cell epitopes*. Drug Discov. Today Biosilico, 2004. **2**(1): p. 18-23.
55. Cohen, M.L., *Changing patterns of infectious disease*. Nature, 2000. **406**(6797): p. 762-7.
56. WHO, *World Health Report*. 2004, World Health Organization.
57. Robinson, H.L. and R.R. Amara, *T cell vaccines for microbial infections*. Nat Med, 2005. **11**(4 Suppl): p. S25-32.
58. Sakakura, K., et al., *Toward the development of multi-epitope p53 cancer vaccines: An in vitro assessment of CD8(+) T cell responses to HLA class I-restricted wild-type sequence p53 peptides*. Clin Immunol, 2007.
59. Wilson, C., et al., *Development of a DNA vaccine designed to induce cytotoxic T lymphocyte responses to multiple conserved epitopes in HIV-1*. J Immunol, 2003. **171**(10): p. 5611-23.
60. Flower, D.R. and I.A. Doytchinova, *Immunoinformatics and the prediction of immunogenicity*. Appl Bioinformatics, 2002. **1**(4): p. 167-176.

61. Rammensee, H.G., *Antigen presentation--recent developments*. Int Arch Allergy Immunol, 1996. **110**(4): p. 299-307.
62. Rammensee, H.G., *Chemistry of peptides associated with MHC class I and class II molecules*. Curr Opin Immunol, 1995. **7**(1): p. 85-96.
63. Brown, J., et al., *Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1*. Nature, 1993. **364**(6432): p. 33-9.
64. Feltkamp, M.C., et al., *Efficient MHC class I-peptide binding is required but does not ensure MHC class I-restricted immunogenicity*. J. Mol. Immunol., 1994. **31**: p. 1391-401.
65. Rognan, D., et al., *Molecular dynamics simulation of MHC-peptide complexes as a tool for predicting potential T cell epitopes*. Biochemistry, 1994. **33**(38): p. 11476-85.
66. Rognan, D., et al., *Rational design of nonnatural peptides as high-affinity ligands for the HLA-B*2705 human leukocyte antigen*. Proc Natl Acad Sci U S A, 1995. **92**(3): p. 753-7.
67. Sezerman, U., S. Vajda, and C. DeLisi, *Free energy mapping of class I MHC molecules and structural determination of bound peptides*. Protein Sci, 1996. **5**(7): p. 1272-81.
68. Ota, N. and D.A. Agard, *Binding mode prediction for a flexible ligand in a flexible pocket using multi-conformation simulated annealing pseudo crystallographic refinement*. J Mol Biol, 2001. **314**(3): p. 607-17.
69. Fagerberg, T., J.C. Cerottini, and O. Michielin, *Structural prediction of peptides bound to MHC class I*. J Mol Biol, 2006. **356**(2): p. 521-46.
70. Jones, D.T., W.R. Taylor, and J.M. Thornton, *A new approach to protein fold recognition*. Nature, 1992. **358**(6381): p. 86-9.
71. Altuvia, Y. and H. Margalit, *A structure-based approach for prediction of MHC-binding peptides*. Methods, 2004. **34**(4): p. 454-9.
72. Schueler-Furman, O., et al., *Structure-based prediction of binding peptides to MHC class I molecules: application to a broad range of MHC alleles*. Protein Sci, 2000. **9**(9): p. 1838-46.
73. Altuvia, Y., O. Schueler, and H. Margalit, *Ranking potential binding peptides to MHC molecules by a computational threading approach*. J Mol Biol, 1995. **249**(2): p. 244-50.
74. Altuvia, Y., et al., *A structure-based algorithm to predict potential binding peptides to MHC molecules with hydrophobic binding pockets*. Hum Immunol, 1997. **58**(1): p. 1-11.
75. Sali, A. and T.L. Blundell, *Comparative protein modeling by satisfaction of spatial restraints*. J. Mol. Biol., 1993. **234**: p. 774-815.
76. Swindells, M.B. and J.M. Thornton, *Structure prediction and modelling*. Curr. Opin. Biotechnol., 1991. **2**: p. 512-519.
77. Logean, A. and D. Rognan, *Recovery of known T-cell epitopes by computational scanning of a viral genome*. J Comput Aided Mol Des, 2002. **16**(4): p. 229-43.
78. Caflisch, A., P. Niederer, and M. Anliker, *Monte Carlo docking of oligopeptides to proteins*. Proteins, 1992. **13**: p. 223-230.
79. Rognan, D., S.L. Laumoeiller, and A. Holm, et al., *Predicting binding affinities of protein ligands from three-dimensional models: application to peptide binding to*

- class I major histocompatibility proteins*. J. Med. Chem., 1999. **42**: p. 4650-4658.
80. Lim, J.S., S. Kim, and H.G. Lee, et al., *Selection of peptides that bind to the HLA-A2.1 molecule by molecular modelling*. Mol. Immunol., 1996. **33**: p. 221-230.
81. Tong, J.C., T.W. Tan, and S. Ranganathan, *Modeling the structure of bound peptide ligands to major histocompatibility complex*. Protein Sci., 2004. **13**: p. 2523-2532.
82. Tong, J.C., et al., *Prediction of HLA-DQ3.2 β Ligands: evidence of multiple registers in class II binding peptides*. Bioinformatics, 2006. **22**: p. 1232-8.
83. Zen, J., H.R. Treutlein, and G.B. Rudy, *Predicting sequences and structures of MHC-binding peptides: a computational combinatorial approach*. J Comput Aided Mol Des, 2001. **15**(6): p. 573-86.
84. Doytchinova, I.A. and D.R. Flower, *Toward the quantitative prediction of T-cell epitopes: coMFA and coMSIA studies of peptides with affinity for the class I MHC molecule HLA-A*0201*. J Med Chem, 2001. **44**(22): p. 3572-81.
85. Doytchinova, I.A., M.J. Blythe, and D.R. Flower, *Additive method for the prediction of protein-peptide binding affinity. Application to the MHC class I molecule HLA-A*0201*. J Proteome Res, 2002. **1**(3): p. 263-72.
86. Guan, P., et al., *MHCPred: a server for quantitative prediction of peptide-MHC binding*. Nucleic Acids Research, 2003. **31**(13): p. 3621-3624.
87. Guan, P., et al., *MHCPred 2.0: an updated quantitative T-cell epitope prediction server*. Appl Bioinformatics, 2006. **5**(1): p. 55-61.
88. Bui, H.H., et al., *Structural prediction of peptides binding to MHC class I molecules*. Proteins, 2006. **63**(1): p. 43-52.
89. Falk, K., et al., *Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules*. Nature, 1991. **351**(6324): p. 290-296.
90. D'Amaro, J., et al., *A computer program for predicting possible cytotoxic T lymphocyte epitopes based on HLA class I peptide binding motifs*. Hum Immunol, 1995. **43**: p. 13.
91. DiBrino, M., et al., *Identification of the peptide binding motif for HLA-B44, one of the most common HLA-B alleles in the Caucasian population*. Biochemistry, 1995. **34**(32): p. 10130-8.
92. Sette, A., J. Sidney, and C. Oseroff, *HLA DR4w4-binding motifs illustrate the biochemical basis of degeneracy and specificity in peptide-DR interactions*. J. Immunol., 1993. **151**: p. 3163-3170.
93. Mata, M., et al., *The MHC Class I-Restricted Immune Response to HIV-gag in BALB/c Mice Selects a Single Epitope That Does Not Have a Predictable MHC-Binding Motif and Binds to K^d Through Interactions Between a Glutamine at P3 and Pocket D^I*. J Immunol, 1998. **161**: p. 2985-2993.
94. Quesnel, A., et al., *Efficient binding to the MHC class I K(d) molecule of synthetic peptides in which the anchoring position 2 does not fit the consensus motif*. FEBS Lett., 1996. **387**: p. 42-46.
95. Gulukota, K., et al., *Two complementary methods for predicting peptides binding major histocompatibility complex molecules*. J. Mol. Biol., 1997. **267**: p. 1258-1267.
96. Hammer, J., et al., *Promiscuous and allele-specific anchors in HLA-DR-binding*

- peptides. *Cell*, 1993. **74**(1): p. 197-203.
97. Parker, K.C., M.A. Bednarek, and J.E. Coligan, *Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains*. *J Immunol*, 1994. **152**(1): p. 163-175.
 98. Jung, G., et al., *From combinatorial libraries to MHC ligand motifs, T-cell superagonists and antagonists*. *Biologicals*, 2001. **29**(3-4): p. 179-81.
 99. Rajapakse, M., et al., *PRED^{NOD}, a prediction server for peptide binding to the H-2^{g7} haplotype of the non-obese diabetic mouse*. *Autoimmunity*, 2006. **39**(8): p. 645-50.
 100. Reche, P.A., J.P. Glutting, and E.L. Reinherz, *Prediction of MHC class I binding peptides using profile motifs*. *Hum Immunol*, 2002. **63**: p. 701-709.
 101. Reche, P.A., et al., *Enhancement to the RANKPEP resource for the prediction of peptide binding to MHC molecules using profiles*. *Immunogenetics*, 2004. **56**(6): p. 405-19.
 102. Peters, B., et al., *Examining the Independent Binding Assumption for Binding of Peptide Epitopes to MHC-I Molecules*. *Bioinformatics*, 2003. **19**(14): p. 1765-1772.
 103. Peters, B. and A. Sette, *Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method*. *BMC Bioinformatics*, 2005. **6**: p. 132.
 104. Brusica, V., et al., *Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network*. *Bioinformatics*, 1998. **14**(2): p. 121-130.
 105. Honeyman, M.C., et al., *Neural network-based prediction of candidate T-cell epitopes*. *Nat Biotechnol.*, 1998. **16**(10): p. 966-969.
 106. Borrás-Cuesta, F., et al., *Specific and general HLA-DR binding motifs: comparison of algorithms*. *Hum Immunol*, 2000. **61**(3): p. 266-78.
 107. Buus, S., et al., *Sensitive quantitative predictions of peptide-MHC binding by a 'Query by Committee' artificial neural network approach*. *Tissue Antigens*, 2003. **62**: p. 378-84.
 108. Nielsen, M., et al., *Reliable prediction of T-cell epitopes using neural networks with novel sequence representations*. *Protein Sci*, 2003. **12**(5): p. 1007-17.
 109. Baum, L.E. and T. Petrie, *Statistical inference for probabilistic functions of finite state Markov Chains*. *Ann. Math. Stat.*, 1966. **37**: p. 1554-1563.
 110. Mamitsuka, H., *Predicting peptides that bind to MHC molecules using supervised learning of hidden Markov models*. *Proteins*, 1998. **33**(4): p. 460-474.
 111. Duda, R.O., P.E. Hart, and D.G. Stork, *Pattern Classification*. 2nd ed. 2001: Wiley-Interscience.
 112. Savoie, C.J., et al., *Use of BONSAI decision trees for the identification of potential MHC class I peptide epitope motifs*. *Pac Symp Biocomput*, 1999: p. 182-9.
 113. Segal, M.R., M.P. Cummings, and A.E. Hubbard, *Relating amino acid sequence to phenotype: analysis of peptide-binding data*. *Biometrics*, 2001. **57**: p. 632-642.
 114. Vapnik, V.N., *The nature of statistical learning theory*. 1998: Wiley New York.
 115. Donnes, P. and A. Elofsson, *Prediction of MHC class I binding peptides, using SVMHC*. *BMC Bioinformatics*, 2002. **3**(1): p. 25-38.
 116. Zhao, Y., et al., *Application of support vector machines for T-cell epitopes*

- prediction*. Bioinformatics, 2003. **19**(15): p. 1978-1984.
117. Cui, J., et al., *MHC-BPS: MHC-binder prediction server for identifying peptides of flexible lengths from sequence-derived physicochemical properties*. Immunogenetics, 2006. **58**(8): p. 607-13.
 118. Liu, W., et al., *Quantitative prediction of mouse class I MHC peptide binding affinity using support vector machine regression (SVR) models*. BMC Bioinformatics, 2006. **7**: p. 182.
 119. Yang, Z.R. and F.C. Johnson, *Prediction of T-cell epitopes using biosupport vector machines*. J Chem Inf Model, 2005. **45**(5): p. 1424-8.
 120. Mallios, R.R., *Predicting class II MHC/peptide multi-level binding with an iterative stepwise discriminant analysis meta-algorithm*. Bioinformatics, 2001. **17**(10): p. 942-8.
 121. Lin, Z., et al., *Toward the quantitative prediction of T-cell epitopes: QSAR studies on peptides having affinity with the class I MHC molecular HLA-A*0201*. J Comput Biol, 2004. **11**(4): p. 683-94.
 122. Xiao, Y. and M.R. Segal, *Prediction of genomewide conserved epitope profiles of HIV-1: classifier choice and peptide representation*. Stat Appl Genet Mol Biol, 2005. **4**(1): p. Article25.
 123. Bhasin, M. and G.P. Raghava, *Prediction of CTL epitopes using QM, SVM and ANN techniques*. Vaccine, 2004. **22**(23-24): p. 3195-204.
 124. Hakenberg, J., et al., *MAPPP: MHC class I antigenic peptide processing prediction*. Appl Bioinformatics, 2003. **2**(3): p. 155-8.
 125. Kesmir, C., et al., *Prediction of proteasome cleavage motifs by neural networks*. Protein Eng, 2002. **15**(4): p. 287-96.
 126. Nussbaum, A.K., et al., *PAPProC: a prediction algorithm for proteasomal cleavages available on the WWW*. Immunogenetics, 2001. **53**(2): p. 87-94.
 127. Bhasin, M. and G.P. Raghava, *Pcleavage: an SVM based method for prediction of constitutive proteasome and immunoproteasome cleavage sites in antigenic sequences*. Nucleic Acids Res, 2005. **33**(Web Server issue): p. W202-7.
 128. Bhasin, M. and G. Raghava, *Analysis and prediction of affinity of TAP binding peptides using cascade SVM*. Protein Sci., 2004. **13**(3): p. 596-607.
 129. Jojic, N., et al., *Learning MHC I-peptide binding*. Bioinformatics, 2006. **22**(14): p. e227-35.
 130. Deluca, D.S., B. Khattab, and R. Blasczyk, *A modular concept of HLA for comprehensive peptide binding prediction*. Immunogenetics, 2007. **59**(1): p. 25-35.
 131. Singh, H. and G.P. Raghava, *ProPred: prediction of HLA-DR binding sites*. Bioinformatics, 2001. **17**(12): p. 1236-1237.
 132. Wan, J., et al., *SVRMHC prediction server for MHC-binding peptides*. BMC Bioinformatics, 2006. **7**: p. 463.
 133. Tenzer, S., et al., *Modeling the MHC class I pathway by combining predictions of proteasomal cleavage, TAP transport and MHC class I binding*. Cell Mol Life Sci, 2005. **62**(9): p. 1025-37.
 134. Larsen, M., 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**(8): p. 2295-303.

173

135. Reche, P.A. and E.L. Reinherz, *PEPVAC: a web server for multi-epitope vaccine development based on the prediction of supertypic MHC ligands*. Nucleic Acids Res, 2005. **33**(Web Server issue): p. W138-42.
136. Brusic, V., et al., *A neural network model approach to the study of human TAP transporter*. In Silico Biol, 1999. **1**(2): p. 109-21.
137. Doytchinova, I. and D. Flower, *Towards the in silico identification of class II restricted T-cell epitopes: a partial least squares iterative self-consistent algorithm for affinity prediction*. Bioinformatics, 2003. **19**(17): p. 2263-2270.
138. White, G. and W. Seffens, *Using a neural network to backtranslate amino acid sequences*. Electron. J. Biotechnol., 1998. **1**.
139. Henikoff, S. and J.G. Henikoff, *Amino acid substitution matrices from protein blocks*. Proc. Natl Acad. Sci. USA, 1992. **89**: p. 10915-19.
140. Nakai, K., A. Kidera, and M. Kanehisa, *Cluster analysis of amino acid indices for prediction of protein structure and function*. Protein Eng, 1988. **2**(2): p. 93-100.
141. Brusic, V., G. Rudy, and L.C. Harrison, *Prediction of MHC binding peptides using artificial neural networks*, in *Complex systems: Mechanism of adaptation*, R.J. Stonier and X.S. Yu, Editors. 1995, IOS Press: Amsterdam. p. 253?60.
142. Kidera, A., et al., *Statistical analysis of the physical properties of the 20 naturally occurring amino acids*. J. Protein Chem., 1985. **4**: p. 23?5.
143. Milik, M., et al., *Application of an artificial neural network to predict specific class I MHC binding peptide sequences*. Nat Biotechnol, 1998. **16**(8): p. 753-6.
144. Taylor, W.R., *The classification of amino acid conservation*. J Theor Biol, 1986. **119**(2): p. 205-18.
145. Antes, I., S.W. Siu, and T. Lengauer, *DynaPred: a structure and sequence based method for the prediction of MHC class I binding peptide sequences and conformations*. Bioinformatics, 2006. **22**(14): p. e16-24.
146. Doytchinova, I., P. Guan, and D. Flower, *EpiJen: a server for multistep T cell epitope prediction*. BMC Bioinformatics, 2006. **7**: p. 131.
147. Saveanu, L., et al., *Complexity, contradictions, and conundrums: studying post-proteasomal proteolysis in HLA class I antigen presentation*. Immunol Rev, 2005. **207**: p. 42-59.
148. Brusic, V., G. Rudy, and L.C. Harrison, *MHCPEP, a database of MHC-binding peptides: update 1997*. Nucleic Acids Res., 1998. **26**(1): p. 368-371.
149. Rumelhart, D.E., G.E. Hinton, and R.J. Williams, *Learning Representations by Back-Propagating Errors*. Nature, 1986. **323**: p. 533-536.
150. Baum, L.E., *An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes*. Inequalities, 1972. **3**: p. 1-8.
151. Baker, J.K., *The dragon-system - An overview*. IEEE Trans. Acoust. Speech Signal Processing, 1975. **ASSAP-23**(1): p. 24-29.
152. Churchill, G.A., *Stochastic models for heterogeneous DNA sequences*. Bull. Math. Biol., 1989. **51**: p. 79-94.
153. Durbin, R.M., et al., *Biological Sequence analysis*. 1998: Cambridge University Press.
154. Asai, K., S. Hayamizu, and K. Handa, *Predicton of protein secondary structure by the hidden Markvo model*. Comput. Appl. Biosci., 1993. **9**: p. 141-146.

174

155. Rabiner, L.R., *A tutorial on hidden Markov models and selected applications in speech recognition*. Proc. IEEE, 1989. **77**: p. 257-286.
156. Sjolander, K., et al., *Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology*. Comput Appl Biosci., 1996. **12**(4): p. 327-345.
157. Brusic, V., et al., *Prediction of promiscuous peptides that bind HLA class I molecules*. Immunol Cell Biol, 2002. **80**: p. 280-285.
158. Forney, G.D., *The Viterbi algorithm*. Proc. IEEE, 1973. **61**: p. 268-278.
159. Gunn, S.R., *Support Vector Machines for Classification and Regression*. 1998, Faculty of Engineering, Science and Mathematics, School of Electronics and Computer Science, University of Southampton.
160. Fletcher, R., *Practical Methods of Optimization*. 1987: John Wiley and Sons., Inc.
161. Smola, A.J. and B. Scholkopf, *A tutorial on support vector regression*. Statistics and Computing, 2004. **14**(3): p. 199 - 222.
162. Webb, A., *Statistical Pattern Recognition*. 2 ed. 2002: Wiley.
163. Zhang, P., *Model selection via multifold cross validation*. Annals of Statistics, 1993. **21**: p. 299-313.
164. Pagano, M. and K. Gauvreau, *Principles of Biostatistics*. 2nd ed. 2000: Duxbury Thomson Learning.
165. Student, *The probable error of a mean*. Biometrika, 1908. **6**: p. 1-25.
166. Swets, J.A., *Measuring the accuracy of diagnostic systems*. Science, 1988. **240**(4857): p. 1285-1293.
167. Moore, D., *Basic Practice of Statistics*. 4 ed. 2006: WH Freeman Company. 90-114.
168. Sakai, Y., et al., *Vaccination by genetically modified dendritic cells expressing a truncated neu oncogene prevents development of breast cancer in transgenic mice*. Cancer Res., 2004. **64**: p. 8022-8.
169. Buchwald, U.K., et al., *A peptide mimotope of type 8 pneumococcal capsular polysaccharide induces a protective immune response in mice*. Infect. Immun., 2005. **73**: p. 325-33.
170. Udaka, K., et al., *An automated prediction of MHC class I-binding peptides based on positional scanning with peptide libraries*. Immunogenetics, 2000. **51**(10): p. 816-28.
171. Hattotuwigama, C.K., et al., *Quantitative online prediction of peptide binding to the major histocompatibility complex*. J Mol Graph Model, 2004. **22**(3): p. 195-207.
172. Bhasin, M., H. Singh, and G.P. Raghava, *MHCBN: a comprehensive database of MHC binding and non-binding peptides*. Bioinformatics, 2003. **19**(5): p. 665-666.
173. McSparron, H., et al., *JenPep: a novel computational information resource for immunobiology and vaccinology*. J Chem Inf Comput Sci., 2003. **43**: p. 1276-1287.
174. Rammensee, H.G., K. Falk, and O. Rotzschke, *Peptides naturally presented by MHC class I molecules*. Annu Rev Immunol, 1993. **11**: p. 213-44.
175. Scott, C.A., et al., *Crystal structures of two I-Ad-peptide complexes reveal that high affinity can be achieved without large anchor residues*. Immunity, 1998. **8**: p. 319-329.

175

176. Sette, A., et al., *Prediction of major histocompatibility complex binding regions of protein antigens by sequence pattern analysis*. Proc Natl Acad Sci USA, 1989. **86**(9): p. 3296-3300.
177. Gupta, V., et al., *SARS coronavirus nucleocapsid immunodominant T-cell epitope cluster is common to both exogenous recombinant and endogenous DNA-encoded immunogens*. Virology, 2006. **347**(1): p. 127-39.
178. Peters, B., et al., *A community resource benchmarking predictions of peptide binding to MHC-I molecules*. PLoS Comput Biol, 2006. **2**(6): p. e65.
179. Cresswell, P., et al., *Mechanisms of MHC class I-restricted antigen processing and cross-presentation*. Immunol Rev, 2005. **207**: p. 145-57.
180. Abele, R. and R. Tampe, *Function of the transport complex TAP in cellular immune recognition*. Biochim Biophys Acta, 1999. **1461**(2): p. 405-19.
181. Abele, R. and R. Tampe, *Modulation of the antigen transport machinery TAP by friends and enemies*. FEBS Lett., 2006. **580**(4): p. 1156-63.
182. Nijenhuis, M., et al., *Identification of a contact region for peptide on the TAP1 chain of the transporter associated with antigen processing*. J Immunol, 1996. **156**(6): p. 2186-95.
183. van Endert, P., et al., *The peptide-binding motif for the human transporter associated with antigen processing*. J Exp Med, 1995. **182**(6): p. 1883-95.
184. Gadola, S.D., et al., *TAP deficiency syndrome*. Clin Exp Immunol, 2000. **121**(2): p. 173-8.
185. Smith, K.D. and C.T. Lutz, *Peptide-dependent expression of HLA-B7 on antigen processing-deficient T2 cells*. J Immunol, 1996. **156**(10): p. 3755-64.
186. Brusic, V., V.B. Bajic, and N. Petrovsky, *Computational methods for prediction of T-cell epitopes - a framework for modelling, testing, and applications*. Methods, 2004. **34**(4): p. 436-43.
187. Daniel, S., et al., *Relationship between peptide selectivities of human transporters associated with antigen processing and HLA class I molecules*. J Immunol, 1998. **161**(2): p. 617-24.
188. Doytchinova, I., S. Hemsley, and D. Flower, *Transporter associated with antigen processing preselection of peptides binding to the MHC: a bioinformatic evaluation*. J Immunol, 2004. **173**(11): p. 6813-9.
189. Peters, B., et al., *Identifying MHC class I epitopes by predicting the TAP transport efficiency of epitope precursors*. J Immunol, 2003. **171**(4): p. 1741-9.
190. Monji, M., et al., *Identification of a novel human cancer/testis antigen, KM-HN-1, recognized by cellular and humoral immune responses*. Clin Cancer Res, 2004. **10**: p. 6047-57.
191. Neisig, A., et al., *Major differences in transporter associated with antigen presentation (TAP)-dependent translocation of MHC class I-presentable peptides and the effect of flanking sequences*. J Immunol, 1995. **154**(3): p. 1273-9.
192. Groothuis, T.A., et al., *MHC class I alleles and their exploration of the antigen-processing machinery*. Immunol Rev, 2005. **207**: p. 60-76.
193. Lautscham, G., et al., *Processing of a multiple membrane spanning Epstein-Barr virus protein for CD8(+) T cell recognition reveals a proteasome-dependent, transporter associated with antigen processing-independent pathway*. J Exp Med, 2001. **194**(8): p. 1053-68.

176

194. Zimmer, J., et al., *Clinical and immunological aspects of HLA class I deficiency*. *Jm*, 2005. **98**(10): p. 719-27.
195. Buteau, C., S.N. Markovic, and E. Celis, *Challenges in the development of effective peptide vaccines for cancer*. *Mayo Clin Proc*, 2002. **77**(4): p. 339-49.
196. Chelvanayagam, G., *A roadmap for HLA-A, HLA-B, and HLA-C peptide binding specificities*. *Immunogenetics*, 1996. **45**(1): p. 15-26.
197. Zhang, G.L., et al., *Neural Models for predicting viral vaccine targets*. *J Bioinform Comput Biol*, 2005. **3**(5): p. 1207-1225.
198. Southwood, S., et al., *Several common HLA-DR types share largely overlapping peptide binding repertoires*. *J Immunol*, 1998. **160**(7): p. 3363-73.
199. Zhang, G.L., et al., *MULTIPRED: a computational system for prediction of promiscuous HLA binding peptides*. *Nucleic Acids Res*, 2005. **33**: p. W172 - W179.
200. Sidney, J., et al., *Specificity and degeneracy in peptide binding to HLA-B7-like class I molecules*. *J Immunol*, 1996. **157**(8): p. 3480-90.
201. Bachinsky, M.M., et al., *Mapping and binding analysis of peptides derived from the tumor-associated antigen survivin for eight HLA alleles*. *Cancer Immun*, 2005. **5**: p. 6.
202. Benlalam, H., et al., *Identification of five new HLA-B*3501-restricted epitopes derived from common melanoma-associated antigens, spontaneously recognized by tumor-infiltrating lymphocytes*. *J Immunol*, 2003. **171**(11): p. 6283-9.
203. Dorrell, L., et al., *Cytotoxic T lymphocytes recognize structurally diverse, clade-specific and cross-reactive peptides in human immunodeficiency virus type-1 gag through HLA-B53*. *Eur J Immunol*, 2001. **31**(6): p. 1747-56.
204. Kubo, H., et al., *Residue 116 determines the C-terminal anchor residue of HLA-B*3501 and -B*5101 binding peptides but does not explain the general affinity difference*. *Immunogenetics*, 1998. **47**(3): p. 256-63.
205. Schonbach, C., et al., *Fine tuning of peptide binding to HLA-B*3501 molecules by nonanchor residues*. *J Immunol*, 1995. **154**(11): p. 5951-8.
206. Sidney, J., et al., *Several HLA alleles share overlapping peptide specificities*. *J Immunol*, 1995. **154**(1): p. 247-259.
207. Sylvester-Hvid, C., et al., *SARS CTL vaccine candidates; HLA supertype-, genome-wide scanning and biochemical validation*. *Tissue Antigens*, 2004. **63**(5): p. 395-400.
208. Tomiyama, H., et al., *Identification of multiple HIV-1 CTL epitopes presented by HLA-B*5101 molecules*. *Hum Immunol*, 1999. **60**(3): p. 177-86.
209. Ueno, T., H. Tomiyama, and M. Takiguchi, *Single T cell receptor-mediated recognition of an identical HIV-derived peptide presented by multiple HLA class I molecules*. *J Immunol*, 2002. **169**(9): p. 4961-9.
210. Joyce, S. and S.G. Nathenson, *Methods to study peptides associated with MHC class I molecules*. *Curr Opin Immunol*, 1994. **6**(1): p. 24-31.
211. Brusic, V. and J. Zeleznikow, *Computational binding assays of antigenic peptides*. *Lett Pept Sci*, 1999. **6**: p. 313-324.
212. Zhang, G.L., et al., *Prediction of supertype-specific HLA class I binding peptides using support vector machines*. *J Immunol Methods*, 2007. **320**: p. 143-154.
213. Joachims, T., *Making large-Scale SVM Learning Practical*. *Advances in Kernel*

- Methods - Support Vector Learning*. 1999, Cambridge: MIT Press.
214. Kast, W.M., et al., *Role of HLA-A motifs in identification of potential CTL epitopes in human papillomavirus type 16 E6 and E7 proteins*. J Immunol, 1994. **152**(8): p. 3904-12.
 215. Elkington, R., et al., *Ex vivo profiling of CD8+-T-cell responses to human cytomegalovirus reveals broad and multispecific reactivities in healthy virus carriers*. J Virol, 2003. **77**(9): p. 5226-40.
 216. Trivedi, D., et al., *Generation of CMV-specific T lymphocytes using protein-spanning pools of pp65-derived overlapping pentadecapeptides for adoptive immunotherapy*. Blood, 2005. **105**(7): p. 2793-801.
 217. Walboomers, J.M., et al., *Human papillomavirus is a necessary cause of invasive cervical cancer worldwide*. J Pathol, 1999. **189**(1): p. 12-9.
 218. Ruppert, J., et al., *Prominent role of secondary anchor residues in peptide binding to HLA-A2.1 molecules*. Cell, 1993. **74**: p. 929-937.
 219. Cortez-Gonzalez, X., et al., *Immunogenic HLA-B7-restricted peptides of hTRT*. Int Immunol, 2006. **18**(12): p. 1707-18.
 220. Fischer, W., et al., *Polyvalent vaccines for optimal coverage of potential T-cell epitopes in global HIV-1 variants*. Nat Med, 2007. **13**(1): p. 100-106.
 221. Khan, A.M., et al., *Large-scale analysis of antigenic diversity of T-cell epitopes in dengue virus*. BMC Bioinformatics, 2006. **7 Suppl 5**: p. S4.
 222. Berzofsky, J.A., et al., *Construction of peptides encompassing multideterminant clusters of human immunodeficiency virus envelope to induce in vitro T cell responses in mice and humans of multiple MHC types*. J Clin Invest, 1991. **88**(3): p. 876-84.
 223. Brown, S.A., et al., *Clustering of Th cell epitopes on exposed regions of HIV envelope despite defects in antibody activity*. J Immunol, 2003. **171**(8): p. 4140-8.
 224. Surman, S., et al., *Localization of CD4+ T cell epitope hotspots to exposed strands of HIV envelope glycoprotein suggests structural influences on antigen processing*. Proc Natl Acad Sci U S A, 2001. **98**(8): p. 4587-92.
 225. Kim, S.K. and R. DeMars, *Epitope clusters in the major outer membrane protein of Chlamydia trachomatis*. Curr Opin Immunol, 2001. **13**(4): p. 429-36.
 226. Gaschen, B., et al., *Diversity considerations in HIV-1 vaccine selection*. Science, 2002. **296**(5577): p. 2354-60.
 227. Price, G.E., et al., *Viral escape by selection of cytotoxic T cell-resistant variants in influenza A virus pneumonia*. J Exp Med, 2000. **191**(11): p. 1853-67.
 228. Bhasin, M. and G.P. Raghava, *A hybrid approach for predicting promiscuous MHC class I restricted T cell epitopes*. J Biosci, 2007. **32**(1): p. 31-42.
 229. King, R.D., et al., *Is it better to combine predictions?* Protein Eng, 2000. **13**(1): p. 15-9.
 230. Trost, B., M. Bickis, and A. Kusalik, *Strength in numbers: achieving greater accuracy in MHC-I binding prediction by combining the results from multiple prediction tools*. Immunome Res, 2007. **3**: p. 5.
 231. Donnelly, J.J., et al., *Preclinical efficacy of a prototype DNA vaccine: enhanced protection against antigenic drift in influenza virus*. Nat Med, 1995. **1**(6): p. 583-7.
 232. Stern, L., et al., *Crystal structure of the human class II MHC protein HLA-DR1*

- complexed with an influenza virus peptide*. Nature, 1994. **368**(6468): p. 215-221.
233. Tamura, S.i. and M. Tateishi, *Capabilities of a Four-Layered Feedforward Neural Network: Four Layers Versus Three*. 1997. **8**(2): p. 251-255.
234. Srinivasan, K.N., et al., *Predictions of Class I T-cell epitopes: Evidence of presence of immunological hot spots inside antigens*. Bioinformatics, 2004. **20**: p. 11-16.
235. Weissert, R., et al., *High Immunogenicity of Intracellular Myelin Oligodendrocyte Glycoprotein Epitopes*. J Immunol, 2002. **169**(1): p. 548-556.
236. Tongchusak, S., *In silico prediction and immunological validation of candida albicans secretory aspartyl proteinase 2 peptides bound to multiple HLA-DRB1 supertypes*, in Department of Microbiology. 2006, Mahidol University.
237. Stoloff, G.A. and W. Caparros-Wanderley, *Synthetic multi-epitope peptides identified in silico induce protective immunity against multiple influenza serotypes*. Eur J Immunol, 2007.
238. Rammensee, H.G., T. Friede, and S. Stevanovic, *MHC ligands and peptide motifs: first listing*. Immunogenetics, 1995. **41**(4): p. 178-228.
239. Singh, H. and G.P. Raghava, *ProPredI: prediction of promiscuous MHC Class-I binding sites*. Bioinformatics, 2003. **19**(8): p. 1009-14.
240. Tongchusak, S., et al., *CandiVF - Candida albicans virulence factor database*. Int J Pep Res Ther, 2005. **11**(4): p. 271-7.
241. Sbail, H., A. Mehta, and A.S. DeGroot, *Use of T cell epitopes for vaccine development*. Curr Drug Targets Infect Disord, 2001. **1**(3): p. 303-13.
242. Velders, M.P., et al., *Defined flanking spacers and enhanced proteolysis is essential for eradication of established tumors by an epitope string DNA vaccine*. J Immunol, 2001. **166**(9): p. 5366-73.
243. Ben-Yedidia, T. and R. Arnon, *Towards an epitope-based human vaccine for influenza*. Hum Vaccin, 2005. **1**(3): p. 95-101.
244. Stevanovic, S., *Identification of tumour-associated T-cell epitopes for vaccine development*. Nat Rev Cancer, 2002. **2**(7): p. 514-20.
245. Zhang, G.L., et al., *PRED^{BALB/c}: a system for prediction of peptide binding to the H2d molecules, a haplotype of the BALB/c mouse*. Nucleic Acids Res, 2005. **33**: p. W180 - W183.
246. Zhang, G.L., et al., *PRED^{TAP}: a system for prediction of peptide binding to the human transporter associated with antigen processing*. Immunome Res, 2006.
247. Zhang, G.L., et al., *Hotspot Hunter: a computational system for large-scale screening and selection of candidate immunological hotspots in pathogen proteomes*. BMC Bioinformatics, 2008. **9**(Suppl 1): p. S19.

Appendix A User Manual of PRED^{BALB/c}

Users have the option to predict peptides binding to all H2^d molecules, H2^d class I molecules, H2^d class II molecules, or a single H2^d molecule. The default selection on the webpage is “all H2^d” molecules. To perform predictions using PRED^{BALB/c}, user must paste a protein sequence into the textbox and assign a name to the sequence. The sequence must contain between nine and 2000 amino acids. If the prediction is run with input sequence containing symbols other than 20 amino acid codes (spaces and carriage returns are allowed) or the total sequence length is outside 9-2000 amino acids range, an error message will be displayed. The input can either be a contiguous protein sequence or a list of peptides, one per line. The default selection on the webpage is “Protein sequence” (Figure A.1 A), which means the input sequence is treated as a contiguous protein sequence (carriage returns and line breaks will be ignored). The PRED^{BALB/c} input processing program decomposes protein sequence (or the list of peptides) into a series of 9-mer peptides overlapping by eight amino acids. Individual 9-mer peptides are then submitted for prediction. Predicted binding scores for all 9-mers are displayed in the result tables (Figure A.1 B). The 9-mer binding scores are within the range 0-10, the higher the score the higher the probability of peptide being binder. PRED^{BALB/c} has the option for plotting the binding scores of all the overlapping 9-mer peptides as a graph, in which X axis represents the start position of a 9-mer peptide and Y axis represents the binding score of the 9-mer peptide (Figure A.1 C). The user can sort the peptides by their binding scores and choose to view only predicted binders with binding scores above a certain threshold.

When users select the input sequence type to be “a list of peptide sequences”, the input sequences separated by carriage returns or line breaks are treated as different peptides (Figure A.2 A). All overlapping 9-mers in each peptide are submitted for prediction. In the result tables, predicted binding scores are represented by the highest individual 9-mer binding score within the input peptide. The predicted binding scores of individual 9-mers in each peptide in the list are not shown (Figure A.2 B). To display the top scoring 9-mer peptides from each input peptide, the user can use the function “View binding peptides at threshold 9” (Figure A.2 B). In the result page (Figure A.2 C), the 9-mers with binding scores equal to or above the threshold 9 are aligned with the input peptides. The predicted 9-mers are displayed with the names of the H2^d alleles to which the 9-mer binding scores are above the threshold. For example, the first input peptide, YPILPEYLQCVK, has binding scores 9.10, 7.00, 5.74, 7.62 and 8.62 to H2-D^d, -K^d, -L^d, I-A^d, and -E^d, respectively (Figure A.2 B). At threshold 9 (SP level 0.9), there are no 9-mer binders to H2^d class II alleles and the 9-mer, ILPEYLQCV, has the highest binding score to H2-D^d, 9.10. Thus, in Figure A.2 C, this 9-mer is aligned with the input peptide and followed by “D^d”.

A step-by-step screenshot walk through of PRED^{BALB/c} is available at <http://antigen.i2r.a-star.edu.sg/predBalbc/HTML/help.html>.

181

A)

PRED^{BALB/c}

Description

Help

Contact Us

For prediction of MHC binding peptides in a protein, select the MHC alleles of your interest and paste the protein sequence in the textbox.

Predict peptides binding to

☒ All H2^d alleles

☐ H2^d class I alleles

☐ H2^d class II alleles

☐ Single H2^d allele

H-Ed

ENTER YOUR PROTEIN SEQUENCES: (Please use input sequence of less than 2000 amino acids.)
[Click here for the input data format](#)

ATSTKKLHKPEATL IKAIDGDTVKLHYKQPMFRLLLVDTPETKHPKKGVEKYGPEAS
AFTKKMVENAKKIEVEFDKGRTDKYGRGLAYIYADGKMNVEALVRQGLARKVATVYKPN
NTHQHLRKSEAQAQKKEKLNIUSEDNADSGG

☒ Protein sequence

☐ A list of peptide sequences

SEQUENCE NAME:

Nase

Submit

Reset

B)

The Predicted Binding of Sequence **Nase** to All H2^d alleles Using Matrix method

Fri Apr 15 18:09:38 2005

Sort by Prediction Score

View binding peptides at threshold

8

View ?

Prediction to H2-D ^d Plot Prediction Score			Prediction to H2-K ^d Plot Prediction Score			Prediction to H2-L ^d Plot Prediction Score			Prediction to I-A ^d Plot Prediction Score			Prediction to I-E ^d Plot Prediction Score		
Position	Peptide	Prediction	Position	Peptide	Prediction	Position	Peptide	Prediction	Position	Peptide	Prediction	Position	Peptide	Prediction
1	ATSTKQLHK	3.46	1	ATSTKQLHK	0.98	1	ATSTKQLHK	1.56	1	ATSTKQLHK	4.74	1	ATSTKQLHK	
2	TSTKQLHKE	3.52	2	TSTKQLHKE	1.60	2	TSTKQLHKE	6.20	2	TSTKQLHKE	7.30	2	TSTKQLHKE	
3	STKQLHKPE	1.32	3	STKQLHKPE	1.80	3	STKQLHKPE	1.50	3	STKQLHKPE	4.78	3	STKQLHKPE	
4	TKQLHKPEA	2.04	4	TKQLHKPEA	1.30	4	TKQLHKPEA	1.80	4	TKQLHKPEA	9.70	4	TKQLHKPEA	
5	KQLHKPEAT	4.00	5	KQLHKPEAT	2.30	5	KQLHKPEAT	1.34	5	KQLHKPEAT	1.88	5	KQLHKPEAT	
6	KLHKPEATL	6.06	6	KLHKPEATL	9.08	6	KLHKPEATL	6.28	6	KLHKPEATL	1.38	6	KLHKPEATL	
7	LHKPEATLJ	8.54	7	LHKPEATLJ	6.10	7	LHKPEATLJ	1.50	7	LHKPEATLJ	7.56	7	LHKPEATLJ	
8	HKEPATLIK	2.46	8	HKEPATLIK	1.80	8	HKEPATLIK	0.90	8	HKEPATLIK	3.08	8	HKEPATLIK	
9	KEPATLIKA	8.90	9	KEPATLIKA	2.20	9	KEPATLIKA	0.56	9	KEPATLIKA	4.54	9	KEPATLIKA	
10	EPATLIKAI	7.30	10	EPATLIKAI	6.06	10	EPATLIKAI	4.54	10	EPATLIKAI	4.12	10	EPATLIKAI	
11	PATLIKAID	3.02	11	PATLIKAID	1.00	11	PATLIKAID	0.40	11	PATLIKAID	6.70	11	PATLIKAID	
12	ATLIKAIDG	4.00	12	ATLIKAIDG	0.96	12	ATLIKAIDG	2.28	12	ATLIKAIDG	9.20	12	ATLIKAIDG	
13	TLIKAIDGD	2.04	13	TLIKAIDGD	5.36	13	TLIKAIDGD	1.42	13	TLIKAIDGD	2.84	13	TLIKAIDGD	
14	LKAIDGDDT	1.88	14	LKAIDGDDT	3.70	14	LKAIDGDDT	1.20	14	LKAIDGDDT	2.30	14	LKAIDGDDT	
15	IKAIDGDDTV	3.34	15	IKAIDGDDTV	5.22	15	IKAIDGDDTV	0.90	15	IKAIDGDDTV	8.30	15	IKAIDGDDTV	

182

C)

Figure A.1 An example of the output pages of PRED^{BALB/c} when input is a single protein. The input protein sequence is *Staphylococcal* nuclease (Nase), the sequence type chosen is “protein sequence” and the H2^d alleles of interest are all H2^d alleles. A) The input page. B) The main result page. The input sequence is decomposed into overlapping 9-mers for prediction of binding scores to each allele. C) Graphical view of the predicted binding scores to I-E^d molecule.

A)

183

B)

The Predicted Binding of Sequence Ova to All H2^d alleles Using Matrix method

Fri Apr 15 19:07:38 2005

Sort by Prediction Score

View binding peptides at threshold View ?

Prediction to H2-D ^d			Prediction to H2-K ^d			Prediction to H2-L ^d			Prediction to I-A ^d		
Position	Peptide	Prediction Score	Position	Peptide	Prediction Score	Position	Peptide	Prediction Score	Position	Peptide	Prediction Score
1	YPILPEYLQCVK	9.10	1	YPILPEYLQCVK	7.00	1	YPILPEYLQCVK	5.74	1	YPILPEYLQCVK	7.62
2	LSGISSAESLK	9.70	2	LSGISSAESLK	6.20	2	LSGISSAESLK	6.90	2	LSGISSAESLK	9.40
3	SSSANLSQISS	6.98	3	SSSANLSQISS	5.62	3	SSSANLSQISS	5.74	3	SSSANLSQISS	8.50
4	ISQAVHAAHAEINEAGR	8.62	4	ISQAVHAAHAEINEAGR	5.30	4	ISQAVHAAHAEINEAGR	7.50	4	ISQAVHAAHAEINEAGR	10.00
5	HIATNAVLFFGR	8.02	5	HIATNAVLFFGR	4.40	5	HIATNAVLFFGR	5.10	5	HIATNAVLFFGR	8.96
6	SAESLKISQAV	4.04	6	SAESLKISQAV	4.96	6	SAESLKISQAV	4.00	6	SAESLKISQAV	8.56

C)

PREDICTED BINDING PEPTIDES AT THRESHOLD 9

YPILPEYLQCVK	Dd
ISGISSAESLK	Ad
SSSANLSQISS	Ad
ISQAVHAAHAEINEAGR	Ad
HIATNAVLFFGR	
SAESLKISQAV	

Figure A.2 An example of the output pages of PRED^{BALB/c} when input is a list of peptides. The input is a list of peptides from chicken ovalbumin and the H2^d alleles of interest are all H2^d alleles. A) The input page. B) The main result page. All 9-mers in each peptide are submitted for prediction. The predicted binding scores are represented by the highest individual 9-mer binding score of each input peptide. C) The alignment view of the predicted binding peptides at threshold 9, which indicates the prediction specificity is 0.9.

Appendix B User Manual of PRED^{TAP}

To perform predictions using PRED^{TAP}, the user needs to paste a protein sequence into the textbox and assign a name to the sequence. The sequence must contain between nine and 2000 amino acids. If the prediction is run with input sequence containing symbols other than 20 amino acid codes (spaces and carriage returns are allowed) or the total sequence length is outside 9–2000 amino acids range, an error message will be displayed and predictions will not be produced. The input can either be a contiguous protein sequence (an amino acid sequence, FASTA or PIR format) or a list of peptides, one per line. The default selection on the webpage is "Protein sequence" (Figure B.1 A), which means the input sequence is treated as a contiguous protein sequence (carriage returns and line breaks will be ignored). The PRED^{TAP} input processing program decomposes protein sequence (or the list of peptides) into a series of 9-mer peptides overlapping by eight amino acids. Individual 9-mer peptides are then submitted for prediction. Predicted binding scores for all 9-mers are displayed in the result tables (Figure B.1 B). The 9-mer binding scores are within the range 0–10, the higher the score the higher the probability of peptide being binder. PRED^{TAP} has an option for plotting the binding scores of all the overlapping 9-mer peptides as a graph, in which X axis represents the start position of a 9-mer peptide and Y axis represents the binding score of the 9-mer peptide. The user can sort the peptides by their binding scores and choose to view only predicted binders with binding scores above a certain threshold (Figure B.1 C).

When users select the input sequence type to be "a list of peptide sequences", the input sequences separated by carriage returns or line breaks are treated as different peptides. All

overlapping 9-mers in each peptide are submitted for prediction. In the result tables, predicted binding scores are represented by the highest individual 9-mer binding score within the input peptide. The 9-mer with the highest binding score in each peptide is displayed as "Binding Core" in the result table (Figure B.2 A). The user can sort the peptides by their binding scores (Figure B.2 B).

A)

The screenshot shows the PREDTAP web interface. At the top, the logo "PREDTAP" is displayed in orange. Below the logo is a navigation bar with three tabs: "Description", "Test", and "Contact Us". The "Test" tab is currently selected. The main content area contains the following text:

For prediction of Transporter associated with Antigen Processing (TAP) binding peptides in a protein, select a prediction method and paste the protein sequence in the textbox. ANN stands for Artificial Neural Networks. HMM stands for Hidden Markov Models.

Select Prediction Method:

☒ ANN ☐ HMM

ENTER YOUR PROTEIN SEQUENCES: (Please use input sequence of less than 2000 amino acids.)

[Click here for the input data format](#)

MHQKRTAMFQDPQERPRKLPQLCTELQTTIHDIIILECVYCKQQLRREVYDFAFRDLCI
VYRDGNPYAVCDKCLFYSKISEYRHYCYSLYGTTLQYQYNKPLCDLLIRCINQKPLC
PEEKQRHLDDKKQRFHNIIRGRWTGRMSSCRSSRTRRETQL

☒ Protein sequence ☐ A list of peptide sequences

SEQUENCE NAME:

B)

The Predicted Binding of Sequence **test to TAP Using Artificial Neural Networks method**

Fri Dec 30 14:22:12 2005

Note: In the table, the binding scores range from 0 to 10. The default threshold is 6 for Artificial Neural Networks method, with scores ≥ 6 referring to predicted **TAP binders**. Scores of < 6 refer to predicted non-binders.

Sequence Name : test

Sort by Prediction Score

View binding peptides at threshold

Plot Prediction Score

Position	Peptide	Prediction
1	MHGKRTAMF	2.12
2	HQKRTAMFQ	2.77
3	QKRTAMFQD	1.10
4	KRTAMFQDP	3.48
5	RTAMFQDPQ	1.48
6	TAMFQDPQE	1.38
7	AMFQDPQER	5.19
8	MFQDPQERP	2.06
9	FQDPQERPR	1.69
10	QDPQERPRK	1.50
11	DPQERPRKL	1.67
12	PQERPRKLP	1.24
13	QERPRKLPO	1.63

C)

PREDICTED BINDING PEPTIDES AT THRESHOLD 6 by Artificial Neural Networks METHOD

Below is the user input protein. The parts highlighted are predicted binding regions.

10	20	30	40	50	60	70
MHGKRTAMFQDPQERPRKLPQLCTELQTTIHDIILECVYCK	QQLLRREYVYDFAFRDLCIVYRDGNPYAVC					
DKCLKFYISKISEYRHVCYSLY	GTTLQYQYNNKPLCDLLIRCI	NCQKPLCP	EEKQRHLDKKQRFHNI	RGWT		
GRCM3CCRSSRTRRETQL						

Figure B.1 The examples of the output pages of PRED^{TAP} for a single protein. The sequence type chosen is "protein sequence". A) The input page. B) The main result page. The input sequence is decomposed into overlapping 9-mers for prediction of binding scores to TAP. C) Alignment view of the predicted TAP binding regions in the input protein.

187

A)

The Predicted Binding of Sequence test to TAP Using Artificial Neural Networks method

Fri Dec 30 14:58:51 2005

Note: In the table, the binding scores range from 0 to 10. The default threshold is 6 for Artificial Neural Networks method, with scores ≥ 6 referring to predicted **TAP binders**. Scores of < 6 refer to predicted non-binders.
Sequence Name: test

Sort by Prediction Score

No.	Peptide	Binding Core	Prediction
1	MHQKRTAMFGDPGE	KRTAMFGDP	3.48
2	RPRKLPQLCTELQT	PRKLPQLCT	3.72
3	TIHDIILECVYCKQ	HDIIILECVY	2.71
4	QLLRREVDFAFRD	RREVDFAF	8.92
5	LCIVYRDGNPYAVC	IVYRDGNPY	6.31
6	DKCLKFYISKISEYR	KFYISKISEY	9.45
7	HYCYSLYGTTLEQQ	SLYGTTLEQ	3.90
8	GRCMSCCRSSRTRRETQL	SRTRRETQL	4.65

B)

Results shown in the order of prediction scores

Rank	Position	Peptide	Binding Core	Prediction
1	6	DKCLKFYISKISEYR	KFYISKISEY	9.45
2	4	QLLRREVDFAFRD	RREVDFAF	8.92
3	5	LCIVYRDGNPYAVC	IVYRDGNPY	6.31
4	8	GRCMSCCRSSRTRRETQL	SRTRRETQL	4.65
5	7	HYCYSLYGTTLEQQ	SLYGTTLEQ	3.90
6	2	RPRKLPQLCTELQT	PRKLPQLCT	3.72
7	1	MHQKRTAMFGDPGE	KRTAMFGDP	3.48
8	3	TIHDIILECVYCKQ	HDIIILECVY	2.71

Figure B.2 An example of the output pages of PRED^{TAP} for a list of peptides. A) The main result page. All 9-mers in each peptide were submitted for prediction. The predicted binding scores are represented by the highest individual 9-mer binding score of each input peptide. The 9-mer with the highest binding score in each peptide is displayed as "Binding Core" in the result table. B) The sorted result page.

Appendix C User Manual of MULTIPRED1

The functions provided by MULTIPRED1 include 1) running predictions; 2) model building; 3) prediction accuracy evaluation; 4) identifying consensus predictions among up to three sets of predictions on a same input protein sequence.

To predict peptides binding to a supertype, users must first select “Run prediction”. The required input is the selection of supertype and prediction method (pre-defined ANN, HMM or SVM). Alternatively, users can select a pre-defined model (built by model building function). By selecting the “Submit” button users get to a sequence input page where the required input is a protein sequence and its name. The length of the input sequence must be between nine and 2000 amino acids. If the input sequence contains symbols other than amino acids (space and carriage returns are allowed) or the sequence is outside the length limits, an error message will be displayed. The input can either be a protein sequence or a list of peptides. The default selection on the webpage is “Protein sequence”, which means the input sequence is treated as one single protein sequence and carriage returns are ignored. If users changed the sequence type to “a list of peptide sequences”, then sequences divided by carriage returns are treated as separate peptides. The processing steps and result pages for the two types of inputs are different. The detailed description on processing steps involved when the input sequence is a protein sequence or a list of peptides are available at <http://antigen.i2r.a-star.edu.sg/multipred1/HTML/faq.html#Q3> and <http://antigen.i2r.a-star.edu.sg/multipred1/HTML/faq.html#Q4>. The 9-mer binding scores range from one to nine (Figure C.1 A), with scores 4-9 referring to predicted binders (8 or

9 referring to high, 6 or 7 to moderate, and 4 or 5 to low confidence of peptide binding). Scores of 1-3 refer to predicted non-binders. MULTIPRED1 saves the prediction result and the users may note down the ID number of the saved jobs for the comparison of prediction results generated by different prediction models (Figure C.1 A). In the result table (Figure C.1 A), "Sum" is the sum total of the individual binding scores of a peptide to the MHC proteins, "Score1" is the top 1 "Sum" in a 30-mer window (A 30-mer window comprises 22 consecutive 9-mer peptides). The scheme for identification of antigenic regions restricted by HLA class I supertypes is based on high scoring individual 9-mers within a window of 30 amino acids [234]. "Score2" is the average of the top 2 "Sum" in a 30-mer window. Similarly, "Score3", "Score4" and "Score5" are the average of the top 3, 4, and 5 "Sum" in a 30-mer window, respectively. To show user a clear view of the binding capacity of an input protein, Score1 through Score5 of all 30-mer peptides of the input protein can be displayed as graphs, in which X axis represents the start position of a 30-mer window and Y axis represents Score1(2/3/4/5) of the 30-mer window. For example, in Figure C.1 B, which is the graph of Score4 of the protein HPV E6, the first three 30-mer windows (starting at positions 1, 2 ,or 3) are 36.82 and the next two windows (starting at positions 4 or 5) have scores 39.50. The following 13 30-mer windows (starting at positions 6-19) have scores higher than 42, the recommended threshold for Score4 for HLA-A2 ANN models (Figure C.1 A), indicating a predicted hotspot, which corresponds to an experimentally determined HLA-A2 hotspot in E6 protein [214]. To locate the individual 9-mers with top binding scores in each 30-mer window, the "align" function can be used. Figure C.1 C shows an example of the alignment view of the top four 9-mers in each 30-mer window. The user can also identify

antigenic regions at a certain threshold by using the “Get hotspots” function (Figure C.1 D). The default values on the webpage are the recommended thresholds for Score4.

When users select the input sequence as “a list of peptide sequences”, the input sequences separated by carriage returns or line breaks are treated as different peptides. All overlapping 9-mers in each peptide are submitted for prediction. In the result tables, predicted binding scores are represented by the highest individual binding score of each input peptide. The predicted binding scores of individual 9-mers in each peptide in the list are not shown (Figure C.2 A). To display the input peptides in the order of their binding scores, the user can use the function “Sort the Result”. In the result page (Figure C.2 B), the input peptides are listed in descending order of their binding scores. To display the predicted 9-mer binders from each input peptide, the user can use the function “Alignment View”. In the result page (Figure C.2 C), the 9-mers with binding scores ≥ 4 are aligned with the input peptides. The predicted 9-mer binders are displayed with the names of the HLA alleles which produced binding scores above the selected threshold.

191

A)

THE PREDICTED PEPTIDE BINDING AFFINITY TO HLA classIA2 using Artificial Neural Network Method

Sun Jan 30 18:31:15 2005

Note: In the table, float values which are less than 4 indicate non-binders, values higher than 4 indicate predicted binders. Higher prediction values indicate higher degree of belief that the predicted 9-mer peptides are MHC binders.

Sequence Name : E6
Result ID : 14235

Result ID : 14235

Plot according to Score: 4

*: In the below table, "Sum" is the sum of the scores in a given row, "Score1" is the top 1 sum score in 30-mer region, "Score2" is the average of the top 2 sum score in 30-mer region, "Score3" is the average of the top 3 sum score in 30-mer region, "Score4" is the average of the top 4 sum score in 30-mer region and "Score5" is the average of the top 5 sum score in 30-mer region.

Sort result according to Score: 4 Plot according to Score: 4 30-mer peptides aligned with top 4 9-mer predictions. 30-mer peptides aligned with top 4 9-mer predictions.

Using Score 4 at threshold 42 (Suggested threshold using Score4 is 42)

unsorted 30-mer table

Position	Peptides	Molecules								Sum	Score1	Score2	Score3	Score4	Score5
		A-0201	A-0202	A-0203	A-0204	A-0205	A-0206	A-0207	A-0209						
1	MHQKRTAMF	2.61	3.21	3.26	2.63	3.66	2.80	2.65	2.61	23.43	44.94	42.92	39.05	36.82	35.06
2	HQKRTAMFQ	2.49	2.82	2.84	2.50	3.13	2.58	2.52	2.49	21.37	44.94	42.92	39.05	36.82	35.06
3	QKRTAMFQD	2.54	2.99	3.00	2.55	3.36	2.66	2.61	2.54	22.25	44.94	42.92	39.05	36.82	35.25
4	KRTAMFQDP	2.47	2.75	2.76	2.48	3.01	2.55	2.53	2.47	21.02	44.94	42.92	42.22	39.50	37.62
5	RTAMFQDPO	2.92	3.88	3.92	2.96	4.37	3.28	2.99	2.92	27.24	44.94	42.92	42.22	39.50	37.62
6	TAMFQDPQE	2.46	2.71	2.74	2.47	2.96	2.53	2.46	2.46	20.79	44.94	43.72	42.78	42.29	40.10
7	AMFQDPQER	3.31	4.49	4.52	3.37	4.99	3.75	3.58	3.31	31.32	44.94	43.72	42.78	42.29	40.10

B)

ALIGNMENT VIEW OF THE PREDICTED BINDING TO HLA classIA2 using ANN

Sequence	Sum	Position
MHQKRTAMFQDPQERPKLPQLCTELQTTI		1-30
KLPLQCTEL	44.94	18-26
QLCTELQTT	40.89	21-29
AMFQDPQER	31.32	7-15
ERPKLPQL	30.13	14-22
HQKRTAMFQDPQERPKLPQLCTELQTTI		2-31
KLPLQCTEL	44.94	18-26
QLCTELQTT	40.89	21-29
AMFQDPQER	31.32	7-15
ERPKLPQL	30.13	14-22
QKRTAMFQDPQERPKLPQLCTELQTTIND		3-32
KLPLQCTEL	44.94	18-26
QLCTELQTT	40.89	21-29
AMFQDPQER	31.32	7-15
ERPKLPQL	30.13	14-22
KRTAMFQDPQERPKLPQLCTELQTTINDI		4-33
KLPLQCTEL	44.94	18-26
QLCTELQTT	40.89	21-29
ELQTTINDI	40.84	25-33
AMFQDPQER	31.32	7-15
RTAMFQDPQERPKLPQLCTELQTTINDII		5-34
KLPLQCTEL	44.94	18-26
QLCTELQTT	40.89	21-29
ELQTTINDI	40.84	25-33
AMFQDPQER	31.32	7-15
TAMFQDPQERPKLPQLCTELQTTINDIIL		6-35
KLPLQCTEL	44.94	18-26
QTTINDIIL	42.50	27-35
QLCTELQTT	40.89	21-29
ELQTTINDI	40.84	25-33

C)

Prediction of ES to classIA2 by ANN

192

D)

HLA classIA2 Hotspots predicted by ANN Method at Threshold 42 for score4

*: In the below table, "Sum" is the sum of the scores in a given row, "Score1" is the top 1 sum score in 30-mer region, "Score2" is the average of the top 2 sum score in 30-mer region, "Score3" is the average of the top 3 sum score in 30-mer region, "Score4" is the average of the top 4 sum score in 30-mer region and "Score5" is the average of the top 5 sum score in 30-mer region.

Protein Sequence Name : E6
Protein Sequence Length : 158

Position	Peptides	Score1'	Score2'	Score3'	Score4'	Score5'
6-35'	TAMFQDPQERPRKLPQLCTELQTTIHD IIL	44.94	43.72	42.78	42.29	40.10
7-36'	AMFQDPQERPRKLPQLCTELQTTIHD IILE	44.94	43.72	42.78	42.29	40.10
8-37'	MFQDPQERPRKLPQLCTELQTTIHD IILEC	44.94	43.72	42.78	42.29	40.12
9-38'	FQDPQERPRKLPQLCTELQTTIHD IILECV	44.94	43.72	42.78	42.31	42.01
10-39'	QDPQERPRKLPQLCTELQTTIHD IILECVY	44.94	43.72	42.78	42.31	42.01
11-40'	DPQERPRKLPQLCTELQTTIHD IILECVYC	44.94	43.72	42.78	42.31	42.01
12-41'	PQERPRKLPQLCTELQTTIHD IILECVYCK	44.94	43.72	42.78	42.31	42.01
13-42'	QERPRKLPQLCTELQTTIHD IILECVYCKQ	44.94	43.72	42.78	42.31	42.01
14-43'	ERPRKLPQLCTELQTTIHD IILECVYCKQQ	44.94	43.72	42.78	42.31	42.01
15-44'	RPRKLPQLCTELQTTIHD IILECVYCKQQL	44.94	43.72	42.78	42.31	42.01
16-45'	PRKLPQLCTELQTTIHD IILECVYCKQQLL	44.94	43.72	42.94	42.43	42.12
17-46'	RKLPQLCTELQTTIHD IILECVYCKQQLLR	44.94	43.72	42.94	42.43	42.12
18-47'	KLQLCTELQTTIHD IILECVYCKQQLLR	44.94	43.72	42.94	42.43	42.12

Predicted hotspots sorted by position:

Position	Sequences	score4'	Length
6-47'	TAMFQDPQERPRKLPQLCTELQTTIHD IILECVYCKQQLLR	42.33	42

Figure C.1 An example of the output pages of MULTIPRED1 when input is a single protein sequence. The input protein sequence is HPV type 16 E6, the prediction method used is ANN and the HLA supertype of interest is HLA-A2. A) The main result page. The input sequence is truncated into overlapping 9-mers for prediction of binding scores to multiple HLA-A2 alleles, *0201, *0202, *0203, *0204, *0205, *0206, *0207 and *0209. The red ovals are added by authors for clarity of viewing. B) Example graph of Score4. C) Alignment view of the top 4 9-mers in 30-mer windows. D) Prediction hotspot region is 6-47 at threshold 42.

A)

THE PREDICTED PEPTIDE BINDING AFFINITY TO HLA classIA3 using Artificial Neural Network Method

Mon Feb 14 18:42:56 2005

Note: In the table, float values which are less than 4 indicate non-binders, values higher than 4 indicate predicted binders. Higher prediction values indicate higher degree of belief that the predicted 9-mer peptides are MHC binders.

Sequence Name : HCV
Result ID : 9879

*: In the below table, "Sum" is the sum of the scores in a given row.

Other Display formats of predict result: [Alignment View](#) | [Sort the Result](#) | [Plot Sum Value](#) |

Position	Peptides	Molecules						Sum*	
		A-0301	A-0302	A-1101	A-1102	A-3101	A-3301		A-6801
1	MSTNPKEFRKTKRN	7.13	6.38	6.75	6.75	6.78	6.65	6.24	46.68
2	KEFRKTKRNTLRP	6.84	5.88	6.16	6.16	6.61	6.42	5.55	43.62
3	TKRNTLRPPQDVRF	5.30	2.89	3.33	3.33	4.46	4.04	2.39	25.74
4	TLRRPQDVRFPGGG	5.30	2.89	3.33	3.33	4.46	4.04	2.39	25.74
5	QDVRFPGGGQIVGG	4.23	1.80	2.38	2.38	3.67	2.71	1.54	18.71
6	PGGGQIVGGVYLLP	4.69	1.93	2.58	2.58	4.00	3.10	1.79	20.67
7	QIVGGVYLLPRRGP	5.73	3.24	3.80	3.80	5.02	4.44	2.67	28.70
8	VYLLPRRGPRLGVR	6.21	4.47	4.86	4.86	5.81	5.39	4.00	35.60
9	RRGPRLGVATRK	6.52	4.95	5.46	5.46	5.94	5.63	4.60	38.56
10	RLGVATRKTSERS	6.52	4.95	5.46	5.46	5.94	5.63	4.60	38.56
11	ATRKTSERSQPRGR	7.05	6.06	6.51	6.51	6.63	6.48	5.87	45.11
12	SERSQPRGRRQPIP	5.67	3.76	4.17	4.17	5.09	4.71	3.15	30.72
13	QPRGRRQPIKARQ	5.51	3.40	3.83	3.83	4.60	4.22	2.87	28.26
14	RQPIKARQPEGRA	5.14	2.49	3.00	3.00	4.74	3.90	2.04	24.31

193

B)

THE PREDICTED BINDING TO classA3 using ANN

Peptide Sequence Name : HCV

Sorted Result

Rank	Position	Peptides	Molecules						Sum	
			A-0301	A-0302	A-1101	A-1102	A-3101	A-3301		A-6801
1	1	MSTNPKFEKRTKRN	7.13	6.36	6.75	6.75	6.78	6.65	6.24	46.68
2	11	ATKTKTSERSQPRGR	7.05	6.06	6.51	6.51	6.63	6.48	5.97	45.11
3	2	KEFRKTRKNTLRPP	6.84	5.88	6.16	6.16	6.61	6.42	5.55	43.62
4	10	RLGVRATKTSERS	6.52	4.95	5.46	5.46	5.94	5.63	4.60	38.56
5	9	RRQPRLOVRATKTK	6.52	4.95	5.46	5.46	5.94	5.63	4.60	38.56
6	8	VYLLPRRQPRLOVR	6.21	4.47	4.86	4.86	5.81	5.39	4.00	35.60
7	23	PTDPRRRSRNLQKV	6.29	4.42	4.97	4.97	5.59	5.21	4.01	35.46
8	24	RRRSRNLLQKMDTL	6.29	4.42	4.97	4.97	5.59	5.21	4.01	35.46
9	22	RRPSWGPTDPRRRSR	6.10	4.14	4.57	4.57	5.69	5.19	3.59	33.85
10	17	WAGPYWPLYGNE	6.03	3.66	4.25	4.25	5.70	5.07	3.10	32.06
11	18	YPWPLYGNEGGMWA	5.81	3.60	4.05	4.05	5.60	4.91	3.09	31.11
12	12	SERSQPRRRQPIIP	5.67	3.76	4.17	4.17	5.09	4.71	3.15	30.72
13	7	QIVGVYLLPRROP	5.73	3.24	3.80	3.80	5.02	4.44	2.67	28.70
14	13	QPRGRROPPIKVARQ	5.51	3.40	3.83	3.83	4.60	4.22	2.87	28.26
15	36	LPQCSFIFLLALL	5.71	3.03	3.61	3.61	4.81	4.29	2.47	27.53
16	37	SFIFLLALLSCLT	5.34	3.04	3.51	3.51	4.54	4.16	2.55	26.65
17	20	GMGWAGWLLSPRGS	5.36	2.84	3.38	3.38	5.00	4.24	2.41	26.61
18	16	PEGRTWAGPYWPP	5.19	2.81	3.34	3.34	4.93	4.12	2.40	26.13
19	32	LAHGVRLVEDGVNY	5.59	2.76	3.48	3.48	4.38	3.85	2.46	25.99
20	33	VRVLEDGVNVATGN	5.59	2.76	3.48	3.48	4.38	3.85	2.46	25.99

C)

THE PREDICT RESULT OF PEPTIDE BINDING

MSTNPKFEFRKTRN	
STNPKFEFRK	A-0301 7.13
STNPKFEFRK	A-0302 6.38
STNPKFEFRK	A-1101 6.75
STNPKFEFRK	A-1102 6.75
STNPKFEFRK	A-3101 6.78
STNPKFEFRK	A-3301 6.65
STNPKFEFRK	A-6801 6.24
KEFRKTRKNTLRPP	
KTKNTLRR	A-0301 6.84
KTKNTLRR	A-0302 5.88
KTKNTLRR	A-1101 6.16
KTKNTLRR	A-1102 6.16
KTKNTLRR	A-3101 6.61
KTKNTLRR	A-3301 6.42
KTKNTLRR	A-6801 5.55
TKNTLRRPQVRF	
TLRPPQVVR	A-0301 5.30
TLRPPQVVR	A-3101 4.46
TLRPPQVVR	A-3301 4.04
TLRPPQVVRFPGGG	
TLRPPQVVR	A-0301 5.30
TLRPPQVVR	A-3101 4.46
TLRPPQVVR	A-3301 4.04
QVRFPGGGQIVGG	
VRFPGGGQ	A-0301 4.23

Figure C.2 An example of the output pages of MULTIPRED1 when input is a list of peptides. The input protein peptides are from hepatitis C virus, the prediction method used is ANN and the HLA supertype of interest is HLA-A3. A) The main result page. As can be seen here, the input sequence is truncated into overlapping 9-mers for prediction of binding scores to multiple HLA-A3 alleles, *0301, *0302, *1101, *1102, *3101, *3301 and *6801. B) Input peptides displayed in the descending order of binding scores. C) Alignment view of the predicted 9-mer binders.

If the user has 9-mer peptides with known binding affinities to proteins belonging to HLA-A2, A3 or DR supertypes and wants to build his own prediction models, the user can use the “Model build” function in MULTIPRED1. Only 9-mer peptides can be used as training data. The users have the option to use their data only, or combine their data with the existing MULTIPRED1 data and build the model on the server. Currently, users can expect to train an HMM model within one minute and a SVM model within 15 minutes, while training of ANN models may take up to 50 minutes (depending on the size of the training dataset) – there are actually four ANNs trained in the background. The ANN models trained by the same dataset are usually slightly different because the initial weights of networks are assigned randomly[197]. To make the trained models more stable,

the training is repeated four times, and four sets of weights are trained – the predictions are the averages of these four predictions. When the model building request is submitted, an intermediate page (Figure C.3) will be displayed providing the result URL which can be bookmarked for later model retrieval.

Figure C.3 When the model building request is submitted, an intermediate page will be displayed providing the result URL which can be bookmarked for later model retrieval.

If the user has 9-mer peptides with known binding affinities and would like to evaluate the prediction accuracy of a model with these peptides, the user can use the “Accuracy Evaluation” function of MULTIPRED1. The system predicts the binding affinities of the input 9-mers and calculates AROC of the predictions. For each supertype, there are two built-in prediction models available in MULTIPRED1. Predictions can also be performed by user-built models. MULTIPRED1 may, therefore, produce several sets of predictions for the same sequence. The comparison of predictions helps identify the most promising peptides picked up as predicted binders by multiple models. The comparison is facilitated by the “Consensus predictions” function. The user needs to input the individual result IDs (up to three) to the system. The Result IDs must be predictions of the same protein and to the same HLA supertype, if the Result IDs belong to predictions on different proteins, an error message will be displayed. The user can select the analysis of the top 5% or 10% of the predicted binders. In the output table, the top 5% or 10% predictions are displayed in

195

the descending order of their binding scores. The peptides selected by multiple models are highlighted in blue or red.

Appendix D User Manual of Hotspot Hunter

To predict antigenic regions in a single protein, users need to open the “Single Sequence Query” page of Hotspot Hunter, paste the query protein sequence into the query box, assign a name to it and select the HLA supertype of interest. The input protein sequence should be in the FASTA format. An input sequence should be between 15 amino acids (for HLA-DR supertype) or 30 amino acids (for HLA-A2/A3/B7 supertype) and 2000 amino acids. Input sequences that contain symbols other than amino acids (spaces and carriage returns are allowed) or are of lengths outside the acceptable length range, will not be processed – an error message will be displayed. The sequence pasted in the query box is treated as one single protein sequence (carriage returns are ignored). Figure D.1 shows the analysis of the human myelin-oligodendrocyte glycoprotein (MOG) for HLA-DR hotspots. Experimentally identified HLA-DR restricted peptides in the mature sequence of MOG (NCBI Accession: CAA88109) form two hotspots, 73-96 and 121-216 (numbering is relative to the mature sequence of MOG) [235]. The input interface is shown in Figure D.1 A. Four hotspots, 38-56, 79-93, 121-163 and 174-211, were predicted (Figure D.1 B). Three of the four predicted hotspots corresponded to the two experimentally identified regions with a single false positive, 38-56. The “binding strength” represents the sum of binding scores predicted by ANN and SVM models for each hotspot. The predicted hotspots are displayed in two formats: in ascending order of their positions in the sequence and in descending order of their prediction scores. Users can click on “Plot Binding Strength” on the result page to generate a graph providing a clear overall view of the binding capacity of the input sequence. The X-axis represents the

start position of a peptide (window size of 15-amino acids for DR supertype and 30-amino acids for A2, A3 and B7 superotypes) and Y-axis represents the predicted score of the peptide. Figure D.1 C shows a plot of the binding scores for the MOG protein with each dot (in blue or red color) representing a 15-amino acids peptide. For example, the first 15-amino acid peptide is plotted at position 1 with a score of 63.53. Individual peptides with prediction scores greater than or equal to the threshold 65 for DR supertype are considered as positive binders to multiple DR alleles and are displayed as red dots, while the non-binders below the threshold are displayed as blue dots.

“Multiple Sequences Query” function allows users to predict hotspots in a set of protein sequences by submitting a multiple sequence alignment file either in ALN (ClustalW) or PHY (PHYLP) format (http://www.ebi.ac.uk/help/formats_frame.html). The results page provides a global view of the localizations of the hotspots in the proteome set, enabling analysis of antigenic diversity and shift of hotspots across protein variants. An example of the Hotspot Hunter analysis of a set of dengue virus type 1 capsid sequences is shown in Figure D.2. The users may bookmark the URL of the result page (shown in Figure D.2 B) to access the results which are kept for 24 hours on the server. Alternatively, they can input their email address in the textbox and the result file will be emailed to them once the prediction process is complete. The consensus of the input protein sequences is displayed on top of the result page (Figure D.2 C). The predicted hotspot regions in each protein of the set are highlighted in yellow. Consensus hotspots are defined as hotspots present in at least 50% of the sequences (subject to presence in at least two sequences submitted for prediction) and are displayed at the bottom of the result page. These consensus hotspots are of interest to immunologists as they are shared across

multiple protein variant sequences.

“Target Selection” function for multiple sequence queries employs a computational method to identify candidate peptides for the formulation of an epitope-based vaccine that could be effective against multiple variants of the pathogen and is also applicable to a large proportion of the human population. The predicted peptides from consensus hotspots of aligned protein sequences are analyzed for selection of best targets for further experimental validation. All the predicted consensus hotspots are ranked to aid in the decision support, aiming to select the best consensus hotspot. Figure S4 on Hotspot Hunter FAQ page (<http://antigen.i2r.a-star.edu.sg/hh/HTML/faq.html#Q5>) summarizes the process of decision-making in selecting the best consensus hotspot and best candidate targets. The output of this analysis is a comprehensive report, which facilitates interpretation of results and the selection of validation experiments. A sample of the target selection analysis output for dengue virus type 1 NS2b protein is provided in Figure D.3. Best peptide targets for experimental validation are highlighted in yellow. Antigenically redundant peptides are peptides whose complete 9-mer antigenic diversity is covered by the other peptides at the consensus hotspot position, thus they can be ignored without loss of information on 9-mer antigenic diversity among the peptides [223]. 9-mer antigenic diversity was studied because they represent the predominant length of binding cores of T-cell epitopes [232].

Hotspot Hunter performs predictions on four HLA supertypes, HLA-A2, -A3, -B7 and -DR. It allows integration of the prediction results of the four supertypes for identification of hotspots common across multiple supertypes, which is facilitated by the function “Hotspots Common Across Supertypes”. The user needs to upload the output files

199

(generated by “Multiple Sequences Query”) of the supertypes of interest for the same set of sequences. Hotspots common across any combination of the four supertypes can be analyzed.

A)

Hotspot Hunter
A system for prediction and analysis of T-cell epitope hotspots

Home Single Sequence Query Multiple Sequences Query Target Selection Hotspots Common Across Supertypes FAQ Contact

To predict immunological hotspots inside a protein, paste the protein sequence into the textbox.

Input Protein Sequence: (The length limit of the sequence is 2000 residues.)

[Click here for data format and sample sequences](#)

GQFRVIGPRHPRIALVGDDELPCRIISPGKNATGMEVGVYRPPFSRVVHLYRNGKDQGDQAPETRGRT
LLKDAIGEGKVTLRIRNVRFSDGEGFTCFRDSYQEEAAMELKVEDPFYVWSPGVLLVAVLPVLLQI
TLGLVFLCLQYRLRGKLRRAEIEHLHRTFDPHFLRVPCWKITLFFVIVPVLGPLVALIICYNWLHRRLAGQF
LEELRNPF

Protein Name:

Select HLA Supertype:

B)

200

C)

Figure D.1 An example of the input and output pages of Hotspot Hunter when performing “Single Sequence Query”. The input protein sequence is MOG and the HLA supertype of interest is HLA-DR. A) The input page. B) Prediction result page. C) Plot of predicted scores vs. the amino acid positions in the protein sequence.

A)

Hotspot position	Rank	Predicted peptide sequence	Frequency of peptide sequence in input dataset (%) [#]
8-30	2	IMAVGIVSILLSSLLKNDVPLAG	72.73
		IMAIGIVSILLSSLLKNDVPLAG	22.73
		IMAVGIVSILLSSFLKNDVPLAG	4.55
32-52	1	LIAGGMLIACYVISGSSADLS	95.45
		LIAGGMLIACYVISGTSADLS	4.55
93-120	3	DTLTILLKATLLAVSGVYPMSIPATLFV	36.36
		DTITILLKATLLAISGVYPMSIPATLFV	13.64
		DTLTILLEATLLAVSGVYPLSIPATLFV	4.55
		DTLTILLKATLLAVSGVYPISIPATLFV	9.09
		ATLTILLKATLLAVSGVYPLSVPATLFV	4.55
		DTLTILLKATLLAVSGVYPMSIPATLFL	4.55
		DTLTILLKATLLAISGVYPMSIPATLFV *	18.18
		DTLTILLKATLLAVSGVYPLSIPATLFV *	9.09

Key:

*

Antigenically redundant sequence

Best peptide targets for experimental validation

#

Only considering the predicted peptide sequences (highlighted in yellow) of the consensus hotspot in the input dataset

Figure D.3 An example of the output page of Hotspot Hunter when using the “Target Selection” function. The input is a HLA-DR supertype prediction result page for dengue virus type 1 NS2b sequences in HTML format (output of "Multiple Sequences Query" function).