

*In silico* inference of immunological relationships between protein  
antigens based on their cytotoxic T-lymphocyte epitope  
repertoires

by

Werner Smidt

Submitted as partial fulfillment of the degree

*Magister Scientiae* Bioinformatics

in the Faculty of Natural and Agricultural Sciences

Department of Biochemistry

Bioinformatics and Computational Biology Unit

University of Pretoria

November 15, 2010

## Declaration of Originality

I, Werner Smidt, declare that the thesis/dissertation, which I hereby submit for the degree *Magister Scientiae* at the University of Pretoria has not been previously submitted by me for degree purposes at any other University and I take note that, if the thesis/dissertation is approved, I have to submit the additional copies, as stipulated by the relevant regulations at least six weeks before the following graduation takes place and if I do not comply with the stipulations, the degree will not be conferred upon me.

SIGNATURE .....

DATE .....

## Acknowledgements

- Fourie Joubert for displaying exceptional patience and confidence during the epic saga of completing this project
- The National Bioinformatics Network for project funding
- Morten Nielsen, Vladimir Brusic and Darren Flower for valuable correspondence
- The rest of the postgraduates in our lab for insightful discussions, most of which you are probably unaware of the positive impact it had on my project
- Friends and family for support during the course of this project



## List of Abbreviations



|               |  |
|---------------|--|
| <b>ANN</b>    | Artificial Neural Network                              |
| <b>APP</b>    | Antigen Presentation Pathway                           |
| <b>ATP</b>    | Adenosine Tri-Phosphate                                |
| <b>AUC</b>    | Area Under the Curve                                   |
| <b>BLOSUM</b> | Block Substitution Matrix                              |
| <b>C1APP</b>  | Class-I Restricted Antigen Presentation Pathway        |
| <b>CTL</b>    | Cytotoxic T-lymphocyte                                 |
| <b>EBV</b>    | Epstein-Barr Virus                                     |
| <b>ER</b>     | Endoplasmic Reticulum                                  |
| <b>ERAP</b>   | Endoplasmic Reticulum Aminopeptidase                   |
| <b>FN</b>     | False Negative   |
| <b>FP</b>     | False Positive   |
| <b>FPR</b>    | False Positive Rate                                    |
| <b>HCV</b>    | Hepatitis C Virus                                      |
| <b>HIV</b>    | Human Immunodeficiency Virus                           |
| <b>HLA</b>    | Human Leukocyte Antigen                                |
| <b>Hsp</b>    | Heat-shock Protein                                     |
| <b>IC50</b>   | Inhibitor concentration at 50% inhibition              |
| <b>IFN</b>    | Interferon   |
| $K_d$         | Dissociation Constant                                  |
| <b>LANL</b>   | Los Alamos National Laboratory                         |
| <b>MCC</b>    | Matthew's Correlation Coefficient                      |
| <b>MHC</b>    | Major Histocompatibility Complex                       |
| <b>NCBI</b>   | National Center for Biotechnology Information          |
| <b>nM</b>     | Nanomolar  |
| <b>ODE</b>    | Optimally Defined Epitopes                             |
| <b>PAM</b>    | Point Accepted Mutation                                |
| <b>pMHC</b>   | Complexed Peptide and Major Histocompatibility Complex |
| <b>ROC</b>    | Receiver Operator Characteristic                       |
| <b>SMM</b>    | Stabilized Matrix Method                               |
| <b>SVM</b>    | Support Vector Machine                                 |
| <b>TAP</b>    | Transporter Associated with Antigen Presentation       |
| <b>TCR</b>    | T-Cell Receptor  |
| <b>TN</b>     | True Negative  |
| <b>TP</b>     | True Positive  |
| <b>UPGMA</b>  | Unweighted Pair Group Method with Arithmetic Mean      |
| <b>VLAPP</b>  | Variabled Lengthed TAP Predictor                       |

# Contents

|   |             |
|---|-------------|
| <b>Table of Contents</b>  | <b>viii</b> |
| <b>List of Figures</b>  | <b>x</b>    |
| <b>List of Tables</b>   | <b>1</b>    |
| <b>1 Introduction</b>   | <b>1</b>    |
| 1.1 The Antigen Processing Pathway . . . . .                    | 1           |
| 1.1.1 Fragment Generation . . . . .                             | 2           |
| 1.1.2 Transport to the ER . . . . .                             | 4           |
| 1.1.3 Association with the MHC Class I binding groove . . . . . | 5           |
| 1.1.4 Recognition by the TCR of a CTL . . . . .                 | 7           |
| Epitope Cross-Reactivity . . . . .                              | 7           |
| 1.2 Pathway Prediction Tools . . . . .                          | 8           |
| 1.2.1 Performance Measurements . . . . .                        | 8           |
| 1.2.2 Proteasome Cleavage Site Predictors . . . . .             | 10          |
| ProteaSMM . . . . .   | 11          |
| NetChop 20S and NetChop C2.0 . . . . .                          | 12          |
| Neural Networks . . . . .                                       | 13          |
| 1.2.3 TAP Affinity Prediction . . . . .                         | 14          |
| 1.2.4 MHC Ligand Prediction . . . . .                           | 16          |
| Bimas . . . . .   | 16          |
| NetMHC . . . . .  | 17          |
| 1.3 Modeling the Entire Process . . . . .                       | 18          |
| 1.3.1 Other Available Tools . . . . .                           | 20          |
| 1.4 The other Immunological Responses . . . . .                 | 20          |

|          |  |           |
|----------|--|-----------|
| 1.5      | Problem Statement . . . . .                                  | 21        |
| 1.6      | Aims . . . . .   | 21        |
| <b>2</b> | <b>Development of Fortuna</b>                                | <b>22</b> |
| 2.1      | Pathway Predictions . . . . .                                | 22        |
| 2.1.1    | Proteasomal Cleavage Prediction . . . . .                    | 23        |
|          | Quantitative Calculation of Proteasomal Fragments . . . . .  | 24        |
|          | Implementation of ProteaSMM . . . . .                        | 26        |
| 2.1.2    | Variable Lengthed TAP Predictor . . . . .                    | 26        |
|          | Construction of VLTAPP . . . . .                             | 26        |
|          | Implementation of VLTAPP . . . . .                           | 31        |
| 2.1.3    | MHC Affinity and Immunogenicity . . . . .                    | 31        |
|          | MHC Ligand Affinity . . . . .                                | 31        |
|          | MHC Ligand Immunogenicity . . . . .                          | 32        |
| 2.1.4    | Combining Pathway Predictions . . . . .                      | 32        |
|          | Combining TAP and Proteasomal Cleavage Predictions . . . . . | 33        |
|          | Combining All Predictions . . . . .                          | 35        |
| 2.2      | Self-Epitopes Via BLAST . . . . .                            | 36        |
|          | Setting BLAST+ Parameters . . . . .                          | 37        |
|          | Measurement of Cross-Reactivity . . . . .                    | 37        |
|          | Relating Epitopes and BLAST Hits . . . . .                   | 38        |
| 2.3      | Analysis and Visualisation . . . . .                         | 38        |
| 2.3.1    | MHC Treemap and Density Plots . . . . .                      | 39        |
| 2.3.2    | Entropy and Frequency Analysis . . . . .                     | 42        |
|          | Entropy as a Measurement of Sequence Variability . . . . .   | 42        |
|          | MHC Ligand Frequency Calculation . . . . .                   | 43        |
|          | Correlating Entropy and Frequency . . . . .                  | 44        |
|          | SeqLogos for MHC ligands . . . . .                           | 46        |
| 2.3.3    | Epitope Info . . . . .                                       | 47        |
| 2.3.4    | Cluster Analysis . . . . .                                   | 48        |
|          | Calculating Sequence Distances . . . . .                     | 49        |
|          | Clustering of Data . . . . .                                 | 52        |
|          | Visualising The Clustering . . . . .                         | 56        |
| 2.4      | Conclusion . . . . .   | 61        |



|          |  |           |
|----------|--|-----------|
| <b>3</b> | <b>Implementation of Fortuna</b>                           | <b>63</b> |
| 3.1      | Fortuna as a Web-Based Application . . . . .               | 63        |
| 3.1.1    | Server Side Processes . . . . .                            | 64        |
|          | How Predictions are Handled . . . . .                      | 64        |
|          | Management of Jobs and Storage . . . . .                   | 65        |
|          | Server Side Development . . . . .                          | 66        |
| 3.1.2    | Client Side . . . . .                                      | 67        |
| 3.2      | Example of Interface and use of Fortuna . . . . .          | 67        |
| 3.2.1    | Registration . . . . .                                     | 68        |
| 3.2.2    | Starting a New Job . . . . .                               | 68        |
| 3.2.3    | Overview of the Submitted Job<br>. . . . .                 | 70        |
| 3.2.4    | Example of Cluster Analysis<br>. . . . .                   | 70        |
| 3.2.5    | Comparing Sequences<br>. . . . .                           | 71        |
| 3.3      | Conclusion . . . . .                                       | 73        |
| <b>4</b> | <b>Results</b>   | <b>74</b> |
| 4.1      | Performance of VLTAPP . . . . .                            | 74        |
| 4.1.1    | Performance Measurements . . . . .                         | 75        |
|          | Single Performance Measurements . . . . .                  | 76        |
|          | Fragmented Predictor Performance Plots of VLTAPP . . . . . | 77        |
| 4.1.2    | VLTAPP Weight Analysis . . . . .                           | 82        |
| 4.2      | Analysis of CTL Epitopes . . . . .                         | 85        |
| 4.2.1    | Methods . . . . .  | 85        |
|          | HIV-1 Analysis . . . . .                                   | 85        |
|          | Influenza A Analysis . . . . .                             | 89        |
| 4.2.2    | Epitope Analysis Results . . . . .                         | 91        |
|          | Pathway Prediction Results . . . . .                       | 91        |
|          | HIV Pathway Prediction Results . . . . .                   | 92        |
|          | Influenza Pathway Prediction Results . . . . .             | 97        |
| 4.2.3    | Variants of Epitopes . . . . .                             | 100       |
|          | Variants of HIV CTL Epitopes . . . . .                     | 100       |
|          | Variants of Influenza CTL Epitopes . . . . .               | 105       |
| 4.2.4    | Clustering Results . . . . .                               | 111       |



|   |            |
|---|------------|
| HIV Epitope Clustering Results . . . . .                                    | 111        |
| Influenza Epitope Clustering . . . . .                                      | 114        |
| 4.2.5 Self-Epitope Discovery . . . . .                                      | 115        |
| 4.3 Conclusion . . . . .  | 118        |
| <b>5 Conclusionary Discussion</b>   | <b>119</b> |
| 5.1 Identification of Pitfalls in CTL Epitope Prediction . . . . .          | 119        |
| 5.1.1 Problems with POPI and Immunogenicity Prediction in General . . . . . | 119        |
| 5.1.2 Cross-reactivity . . . . .  | 121        |
| 5.1.3 Proteasomal Cleavage Prediction . . . . .                             | 121        |
| 5.1.4 MHC Prediction . . . . .  | 121        |
| 5.2 Bioinformatics Facilitating CTL Based Vaccine Design . . . . .          | 121        |
| 5.3 The use of Prediction Tools in this Study . . . . .                     | 122        |
| 5.4 Conclusion . . . . .  | 122        |
| <b>6 Summary</b>  | <b>124</b> |
| <b>Bibliography</b>   | <b>124</b> |

## List of Figures

|      |   |    |
|------|---|----|
| 1.1  | Overview of the MHC Class I restricted antigen presentation pathway . . . . .     | 2  |
| 1.2  | The 20S/Constitutive proteasome structure from the side and top . . . . .         | 3  |
| 1.3  | HLA*A0201 associated with a nonamer peptide with the sequence LLFGYPVYV . . . . . | 6  |
| 1.4  | Sensitivity and Specificity plots of an ANN based MHC ligand predictor . . . . .  | 9  |
| 1.5  | Demonstration of a ROC analysis . . . . .   | 10 |
| 1.6  | Example of an SMM calculation . . . . .   | 11 |
| 1.7  | ROC plots of N-terminal weighted and unweighted schemes . . . . .                 | 15 |
| 1.8  | Difference between a Sparsed and BLOSUM encoded ANN . . . . .                     | 19 |
| 2.1  | Fortuna Development Flow Diagram . . . . .  | 23 |
| 2.2  | Calculation of Proteasomal Fragment Probabilities . . . . .                       | 27 |
| 2.3  | Word list generation . . . . .  | 27 |
| 2.4  | MHC Treemap Example . . . . .   | 40 |
| 2.5  | MHC Density Plots Example . . . . .   | 41 |
| 2.6  | Smooth Frequency Example . . . . .  | 45 |
| 2.7  | Entropy vs Frequency Example . . . . .  | 46 |
| 2.8  | SeqLogo Example . . . . .   | 47 |
| 2.9  | MHC ligand variability example . . . . .  | 47 |
| 2.10 | UPGMA Example . . . . .   | 54 |
| 2.11 | Immunological Distance Matrix Construction . . . . .                              | 55 |
| 2.12 | Asymmetric Dendrograms . . . . .  | 56 |
| 2.13 | Heatmap Cluster Example . . . . .   | 58 |
| 2.14 | Grouping of Heatmap Clusters . . . . .  | 59 |
| 2.15 | Heatmap Grouping Annotation . . . . .   | 60 |
| 2.16 | Immunological Comparisons . . . . .   | 62 |



|      |  |     |
|------|--|-----|
| 3.1  | Fortuna Overview . . . . .   | 64  |
| 3.2  | Main Prediction Process Overview . . . . .   | 65  |
| 3.3  | Registration Screen . . . . .  | 68  |
| 3.4  | Creating a new Fortuna Job . . . . .   | 69  |
| 3.5  | Completed Job Overview . . . . .   | 70  |
| 3.6  | Clustering Example . . . . .   | 72  |
| 3.7  | Comparing Sequences Example . . . . .  | 73  |
|      |  |     |
| 4.1  | VLTAPP ROC Curve comparison between different lengthed peptides . . . . .  | 80  |
| 4.2  | FPPP of VLTAPP . . . . .   | 81  |
| 4.3  | Weights of VLTAPP Neural Network Inputs . . . . .  | 83  |
| 4.4  | RNA Levels versus SLYNTVATL variants . . . . .   | 103 |
| 4.5  | Variants of selected epitopes over time. The labels of the Figures indicated the<br>Influenza in which the epitope exists. . . . . | 109 |
| 4.6  | Timeline for HLA*A1101 Epitope Scores . . . . .  | 110 |
| 4.7  | Clustering of HIV Sequences based on HLA*A0201 Epitopes . . . . .  | 113 |
| 4.8  | CTL Epitope Classification of Influenza H1N1 and H3N2 Serotypes . . . . .  | 115 |
| 4.9  | Influenza H1N1 Clustering Heatmap . . . . .  | 116 |
| 4.10 | Influenza H3N2 Clustering Heatmap . . . . .  | 117 |
| 4.11 | Self-Epitope Plot . . . . .  | 118 |

## List of Tables

|      |   |    |
|------|---|----|
| 1.1  | Comparison of important residues in ProteaSMM and NetChop . . . . .                     | 13 |
| 1.2  | Correlation between alpha coefficient and HLA-Allotype . . . . .                        | 16 |
| 2.1  | Dataset size with variable lengthed peptides . . . . .                                  | 28 |
| 2.2  | Adjustment of IC50 values from entries using different standard peptide concentrations. | 29 |
| 2.3  | Amino acid properties used as input parameters . . . . .                                | 30 |
| 2.4  | TAP input from Proteasomal Prediction . . . . .   | 32 |
| 2.5  | Relationship between Scores and Amount . . . . .  | 35 |
| 2.6  | Epitope Counts per Allotype at position in Sequences. . . . .                           | 40 |
| 2.7  | Epitope Counts per Allotype at position in Sequences. . . . .                           | 41 |
| 2.8  | Shannon Entropy Example . . . . .   | 43 |
| 2.9  | MHC Ligand Frequency Calculation. . . . .   | 44 |
| 2.10 | Pathway Prediction Combination Criteria . . . . .                                       | 49 |
| 2.11 | Epitope Scoring Information Table . . . . .   | 50 |
| 2.12 | Epitope Information for Sequences. . . . .  | 51 |
| 2.13 | Calculation of Basic Immunological Distance . . . . .                                   | 52 |
| 2.14 | Calculation of Weighted Immunological Distance . . . . .                                | 52 |
| 3.1  | Steps in Prediction Process. . . . .  | 65 |
| 3.2  | Word List Generation . . . . .  | 66 |
| 3.3  | Programming Package Utilised for Server Side Development. . . . .                       | 66 |
| 4.1  | Performance Measurements on VLTAPP . . . . .  | 75 |
| 4.2  | VLTAPP Single performance measurement for all entries. . . . .                          | 76 |
| 4.3  | VLTAPP Single performance measurement for Ligands longer than nine amino acids. . . . . | 77 |



|      |  |     |
|------|--|-----|
| 4.4  | Comparison of Length Distribution of longer TAP ligands between VLTAPP and Peters' set . . . . .   | 78  |
| 4.5  | Top End FPPP values vs Single Performance Measurements. . . . .  | 79  |
| 4.6  | Influence of Amino Acids on Score . . . . .  | 84  |
| 4.7  | Search Parameters for selecting HIV sequences . . . . .  | 86  |
| 4.8  | HLA Allotype Frequencies and Prediction Choice . . . . .   | 87  |
| 4.9  | Breakdown of Influenza A Sequence Set . . . . .  | 90  |
| 4.10 | Predictions for many HLA Allotypes of superclass A2. The green, orange and maroon numbers indicate predicted IC50 values at or below 500nM, between 500nM and 1000nM and above 1000nM respectively. Peptides marked with + are those for which only a close match was found in the sequences and those marked with * are peptides for which no close match was found. . . . .  | 94  |
| 4.11 | Prediction Results for Optimal HIV CTL Epitopes . . . . .  | 95  |
| 4.12 | Total Predicted HLA A*201 Epitopes for HIV Proteins . . . . .  | 96  |
| 4.13 | Prediction Results for Optimal Flu CTL Epitopes . . . . .  | 98  |
| 4.14 | Top Predicted HLA A*201 Epitopes for Flu Proteins . . . . .  | 99  |
| 4.15 | Variants for certain HLA*0201 restricted optimal epitopes. The epitope sequence from the literature is shown in the first column. The second and third column show the frequency of MHC binders (i.e. all the epitope variants that have sufficient binding affinity to the HLA molecule) and the the averaged entropy of the sequences. Nominal sequence and its variants are shown in column four. Column five and six show the sequence variants in SeqLogo format. . . . . | 102 |
| 4.16 | Entropy/Frequency . . . . .  | 104 |
| 4.17 | Variants for certain HLA*0201 restricted optimal epitopes of Influenza A H1N1. .   | 106 |
| 4.18 | Variants for certain HLA*0201 restricted optimal epitopes of Influenza A H3N2. .   | 107 |
| 4.19 | Variants for certain HLA*0201 restricted predicted epitopes of Influenza A H3N2 with high sequence entropy. . . . .  | 107 |
| 4.20 | CompSeq . . . . .  | 115 |

## Abstract

The importance of Cytotoxic T-Cell (CTL) responses during the course of intracellular infections has received a lot of attention during the past few decades. CTLs respond to epitopes presented by the Major Histocompatibility Complex (MHC) originating from intracellular proteins for which they have an appropriate T-Cell Receptor (TCR) for. This response is crucial for the control of pathogens such as Influenza, Hepatitis, HIV and others by destroying the cell in which the pathogen replicates. Due to the extreme polymorphism of MHC molecules, Computational Immunology techniques have been developed to detect potential MHC ligands and as a consequence, potential CTL epitopes. The polymorphism factor needs to be taken into account especially when concerning the design of vaccines with a CTL response component to maximize population coverage. Tools have been constructed that combine the predictions tools concerning major steps in this pathway, that is, proteasomal cleavage, Transporter associated with Antigen Presentation (TAP) affinity, Major Histocompatibility Complex (MHC) affinity and Immunogenicity. In this study, a novel method is developed to combine the different steps in the pathway, which includes the development of a novel TAP predictor. Furthermore, by using a BLOSUM-based score in conjunction with the epitope prediction results, a novel CTL epitope-based clustering method was developed. Two pathogens with major CTL epitope components, but vastly different mutation rates were chosen to infer whether the aforementioned methods can be used to detect potential CTL epitopes and group sequences together based on shared immunogenicity.

The immune system is a collection of cells and tissues with the main task of keeping and restoring a state of homeostasis that can be disrupted by pathogenic entities. Understanding the complex mechanism by which the immune system works is crucial in this modern age. With the increase in computing power over the last few decades, computational modelling of different processes have been made possible. One of the crucial aspects of the immune system is the Cytotoxic T-lymphocyte response. These cells are in charge of destroying cells presenting foreign peptides that could be of pathogenic or cancerous origin (Shevach, 2002, Shedlock and Shen, 2003). Experimental screening of peptides for CTL epitopes is an arduous task, not because of the procedures, but because of the sheer amount of peptides and other agents that need to be screened. Here we will be discussing known aspects of the MHC Class I antigen presentation pathway and the different ways it can be modelled from available experimental data. From scoring matrices to learning techniques like Support Vector Machines and Artificial Neural Networks have been applied in an attempt to solve this problem. We will also see that in designing these tools, certain hypotheses can be formulated from results obtained. An overview of the biological process will be given and with each of the main steps and the corresponding prediction tools will be discussed.

## 1.1 The Antigen Processing Pathway

To understand the problems associated with designing tools for modelling the antigen presentation, it is essential to understand the underlying mechanisms of the pathway. One can imagine the pathway as being a byproduct of protein recycling inside the cell. Proteins have a limited lifespan inside a cell and during their breakdown, some of the peptide fragments might enter the antigen presentation pathway. The pathway can be broken down into three parts:

1. Fragment generation - *How the cell generates a diversity of internal peptides for MHC Class I presentation*
2. Transport to the ER - *How the fragments are transported to the ER*

3. Association with the MHC Class I binding groove - *What determines whether a peptide would bind efficiently to the MHC molecule*

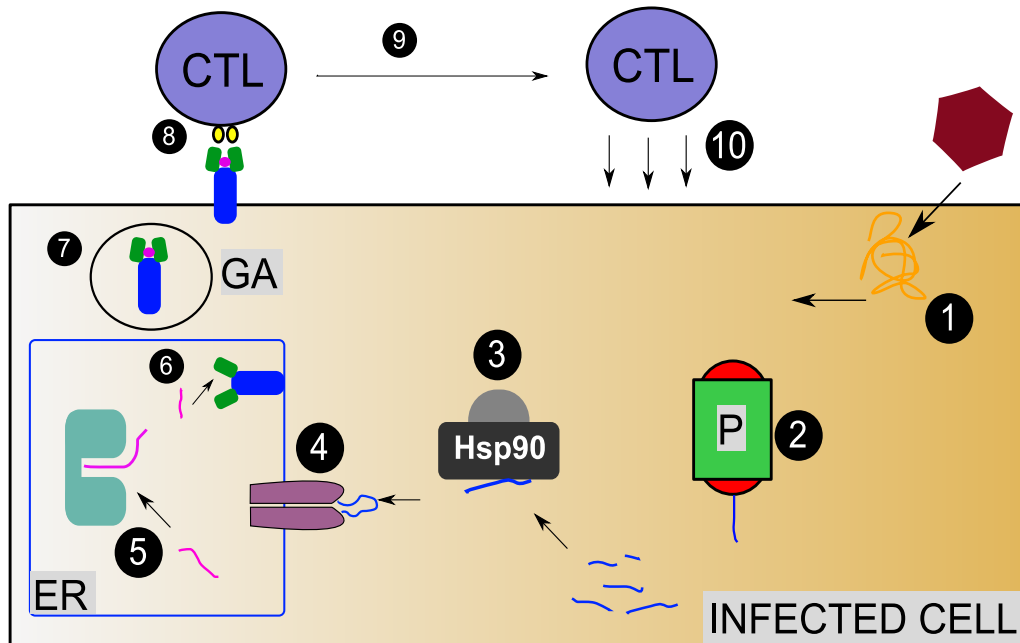


Figure 1.1: 1) A hypothetical virus enters the cell and production of the protein products take place 2) Some of the viral peptides are digested by the cell's proteasomes 3) Fragments are carried by chaperones to the Transporter associated with Antigen Presentation (TAP) 4) TAP transports the peptide into the ER lumen 5) The peptide gets trimmed on its N-terminal by ERAP 6) The peptide binds with MHC 7) The peptide-MHC complex is transported by the Golgi apparatus to the cell membrane 8) A Cytotoxic T-lymphocyte (CTL) with a complimentary receptor binds to the MHC-peptide complex 9) Proliferation and activation of the CTL occurs 10) The CTL signals the cell's destruction. Adapted from (Pamer and Cresswell, 1998, Cresswell *et al.*, 2005)

Recognition by a CD8<sup>+</sup> T-Cell could be considered as a fourth step. With each of the steps there exists a level of redundancy that will be investigated. Starting off with fragment generation, all the consecutive steps and tools designed to model them will be discussed with emphasis on MHC binding and tools associated with it. A rudimentary illustration is shown in Figure 1.1.

### 1.1.1 Fragment Generation

The human proteasome is a 20S barrel shaped structure (see Figure 1.2) with the main purpose of specifically degrading proteins in the cytosol (Unno *et al.*, 2002). It has affinity for proteins marked for degradation, e.g. by ubiquitination. The proteasome's core cleaving regions are isolated from the cytosol, which protects normal cellular proteins from random degradation. Two types of proteasomes can exist within a cell, the constitutive and the IFN $\gamma$  induced immunoproteasome (Pamer and Cresswell, 1998).

They are similar except for three subunits. Under stimulation by IFN- $\gamma$ , subunits X, Y and Z of the constitutive proteasome are replaced by LMP7, LMP2 and MECL-1 respectively (Eleuteri

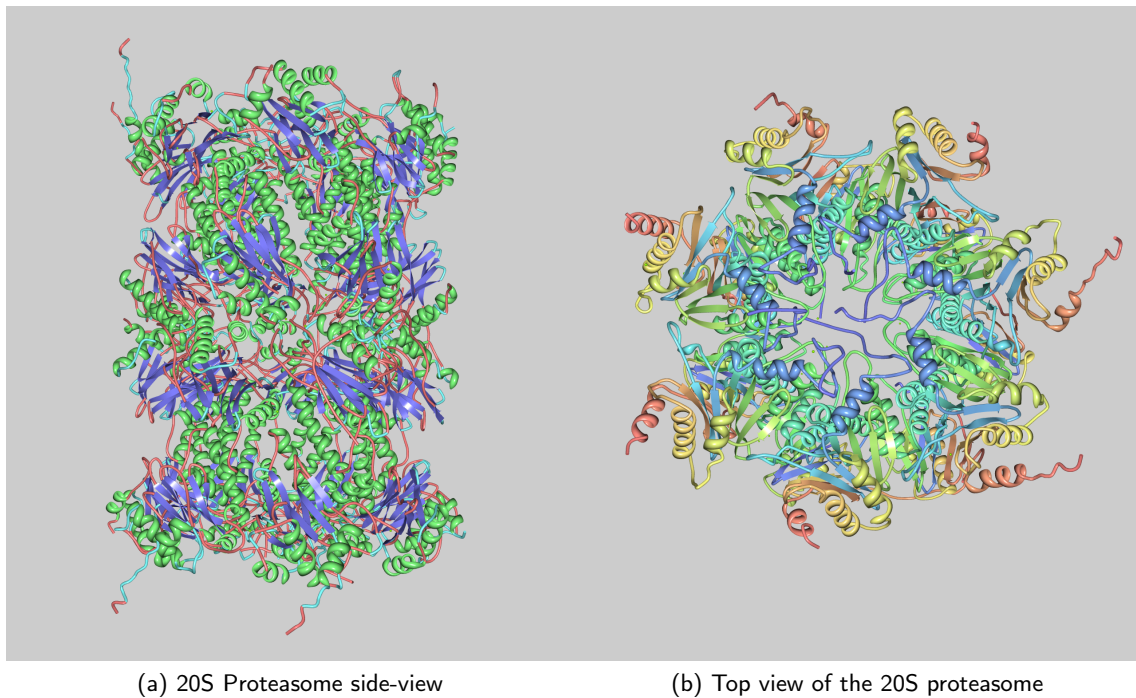


Figure 1.2: The 20S/Constitutive proteasome structure from the side and top. Note, to ease viewing, the top view only shows half of the structure [PDB: 1IRU]. (Unno *et al.*, 2002)

*et al.*, 1997). The constitutive proteasome is more attuned to cutting after acidic residues while the immunoproteasome prefers to cut after hydrophobic residues (Gaczynska *et al.*, 1996). The advantages of altering the specificity of the proteasome has to do with the diversity of peptides generated. Fragments produced by the immunoproteasome usually contain hydrophobic residues at the C-terminal side, which is preferable for MHC Class I binding. There seems to be no further trimming on the C-terminal side, but if the peptide is extended on the N-terminus, the ER contains aminopeptidases (e.g. ERAP) that cut the peptide to an appropriate size (Chang *et al.*, 2005). This does not make the constitutive proteasome useless, since it has been shown that antigen presentation can still occur in the absence of immunoprotease subunits (Pamer and Cresswell, 1998). Also, in the rabbit, skeletal muscle cells do not produce substitution units when stimulated by IFN- $\gamma$ , but are still able to present MHC complexes (Kisselev *et al.*, 1999). Though, one can imagine the total scope of presentable antigens is limited by reduced coverage of cleavage sites in the absence of these substitution units.

A denatured/ubiquitinated protein enters the proteasome and if appropriate residues in the protein are in close proximity to the hydrolases, the protein is cleaved and the peptides exit the proteasome (Kisselev *et al.*, 1999). It has been shown that efficient fragment generation is correlated with the presentation of CTL epitopes (Sijts *et al.*, 1997). Thus peptide fragments generated from other means, such as aborted translation products should prove to be insufficient for adequate peptide presentation.

Though beyond the scope of this review, it is worth mentioning that other factors also

attach to the proteasome enhancing cleavage of the peptides. The PA28 subunits bind to both ends of the proteasome (forming a 26S particle) and enhances cleavage of peptides by changing the conformation of the input peptide, making more cleavage sites accessible at the same time (Kisselev *et al.*, 1999). However, it has been shown that the 26S proteasome complex can produce more products that are smaller than the appropriate length for MHC binding (Kisselev *et al.*, 1999, Fahnestock *et al.*, 1994). Since the proteasome can cleave peptides in an ATP dependent manner, association of subunits such as PA700, an ATPase, can dramatically increase the efficiency of fragment generation.

Two tools, NetChop C2.0 (Kesmir *et al.*, 2002) and ProteaSMM (Tenzer *et al.*, 2005) that predict potential proteasomal cleavage sites will be discussed in Section 1.2.2.

### 1.1.2 Transport to the ER

The MHC complex needs to associate with a potential epitope in the ER. Transport of the proteasomal fragments to the ER is thus an essential step in antigen presentation. Two questions need to be addressed here: *i) How are peptides transported across the membrane of the ER and ii) How are peptides transported to this cross-membrane transporter.* The Transporter Associated with antigen Presentation (hereafter, TAP) facilitates transport of peptides from the cytosol to the lumen of the ER and various chaperones facilitate the transport of peptides to TAP (Wright *et al.*, 2004). The TAP dimer consists of two monomers, TAP.1 and TAP.2. Across mammalian species there exists a lot of similarities between the TAP transporter. However, they do sometimes differ in promiscuity. For instance, the mouse TAP molecule has a preference for hydrophobic residues at the C-terminus of the peptide ligand, whereas the human TAP molecule can efficiently bind peptides with a basic residue at the C-terminal end (Uebel *et al.*, 1997). TAP concerns itself with the terminal residues at either end of a potential ligand. Variation in the terminal residues can significantly influence binding affinity to TAP (van Endert *et al.*, 1995). Identifying the preference for these residues is crucial in designing TAP predictors. The problem with this is that TAP need not bind peptides of lengths appropriate for the MHC Class I binding groove, and therefore exists no "fixed" window in defining a TAP binder (Peters *et al.*, 2003). There is some correlation between the terminal end of a binding peptide and the terminal end for an MHC Class I binding motif (Khan *et al.*, 2000). Still, TAP is able to transport peptides that do not meet these requirements. Chaperones, especially Hsp70 and Hsp90 play important roles in protecting a peptide product of the proteasome from other cytosolic endopeptidases (Callahan *et al.*, 2008).

### 1.1.3 Association with the MHC Class I binding groove

The MHC Class I molecule is probably the most important molecule in antigen presentation as it actually presents a peptide for screening by CTLs. The binding site for the peptide consists of a bed of anti-parallel  $\beta$ -strands, overlaid in a flanked fashion by two anti-parallel  $\alpha$ -helices. The peptide ligand binds in the groove formed by the alpha helices. The heavy chain of the MHC molecule also associates with a  $\beta$ 2-microglobulin (Khan *et al.*, 2000). Assembly of the final MHC- $\beta$ 2M-peptide complex involves other peripheral proteins, and the process can be summarised as (Cresswell *et al.*, 2005):

1. ERp56 associates with tapasin, which associates with TAP.1
2. MHC Heavy chain dissociates from the chaperone, Calnexin and associates with the chaperone Calreticulin and  $\beta$ 2-M associates with the MHC heavy chain.
3. The MHC-Calreticulin- $\beta$ 2-M complex associates with tapasin and ERp57
4. Transported peptides are cleaved by ERAP to appropriate length and if they bind to the MHC binding groove, the MHC- $\beta$ 2-M-peptide complex dissociates and transported to the cell surface by the Golgi apparatus

The article by Cresswell *et al.* (2005) provides an illustration explaining this process. Peptides may also be trimmed by ERAP. ERAP is an IFN- $\gamma$  inducible, ER-associated aminopeptidase and is quite interesting in that it does not cut peptides to a length smaller than 8 amino acids. Only about 30% of the proteasomal fragments can potentially be MHC ligands and 50% of those are too long (Chang *et al.*, 2005). ERAP therefore directly increases the amount of potential MHC ligands. MHC molecules that have no peptide bound in the groove are eventually degraded, but not presented. The author imagines two reasons for this, one being that "empty" MHC molecules would take up space on the cell surface, limiting the visibility of peptide-carrying MHC complexes. The other would be wasting valuable cellular resources on a useless process.

There are six main types of MHC Class I molecules in humans, HLA-A to HLA-F. HLA-A, HLA-B and HLA-C conform to the classical mode of presentation, presenting unmodified peptides and are designated Class Ia molecules. HLA-D, HLA-E and HLA-F are involved in presenting a completely different set of peptides, in terms of modification and length. They are designated Class Ib molecules (Pamer and Cresswell, 1998). Here, we will be focussing on the Class Ia types since the prediction tools discussed later only concerns them.

A very brief overview of the binding process will be discussed here, as not to reiterate when discussing the prediction tools. The MHC binding groove contains position-specific binding pockets that have strong affinities for a specific range of residues. During motif discovery experiments,



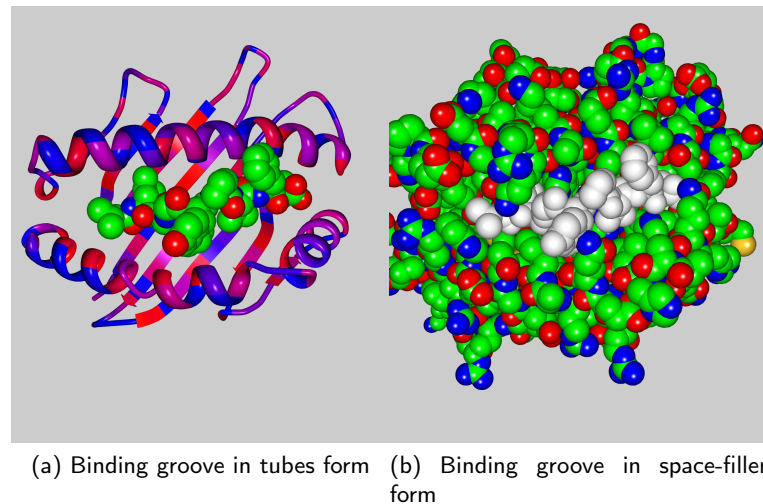


Figure 1.3: HLA\*A0201 associated with a nonamer peptide with the sequence LLFGYPVYV. From left to right, the peptide ligand is arranged in an N-C fashion [PDB: 1DUZ] (Khan *et al.*, 2000).

it was revealed that there exists 2-3 pockets that bind a limited range of amino acids (Rammensee *et al.*, 1995). These positions are called anchor positions and it is essential that the correct amino acids exist in the correct position in order for the potential ligand to bind strongly to the groove.

The binding groove of the *HLA A\*0201* molecule with bound peptide LLFGYPVYV is shown in Figure 1.3. From the clefts on either side of the binding groove, it is easy to see that the length of the binding peptide is limited. However, the peptide still needs to span a majority of the groove to form interactions with crucial binding pockets which exist near the clefts. It has been demonstrated that when the P1 residue of the peptide is removed, binding still occurred, but at a severely lowered affinity (Khan *et al.*, 2000). A stable binding of an MHC ligand changes the conformation of the groove from an open to a closed state. This allows the MHC-peptide complex to have a long half-life, sometimes even tens of hours (Khan *et al.*, 2000).

Even though the anchor residues of a potential ligand is important, a peptide containing them does not by default qualify as an appropriate binder. Other interactions are also important, using the nonamer from before as an example. Correct conformation of the ligand backbone is crucial for orienting residues towards potential binding pockets. In addition to this, various water bridges can form between backbone -CO and -NH groups. Residues might influence the orientation of their neighbours and it is this property that distinguishes two MHC ligand predictors, Bimas and NetMHC (Parker *et al.*, 1994, Nielsen *et al.*, 2004a). The former assumes that contribution of each residue in the potential ligand is independent, while the latter addresses the interdependence problem.

A stable MHC-peptide complex is then transported via the Golgi apparatus to the cell surface where it comes under CTL scrutiny.



### 1.1.4 Recognition by the TCR of a CTL

An appropriate TCR can bind to the MHC-peptide complex and initiate a series of events that would eventually lead to CTL expansion, destruction of the presenting cell. If CD4<sup>+</sup> T-Cells are stimulated via MHC Class II, they retain long lasting memory of this epitope (Shedlock and Shen, 2003). The immunogenicity, or ability to induce an immune response, of a peptide has a low correlation with its binding affinity. Factors that may influence immunogenicity are (Ochoa-Garay *et al.*, 1997):

- Affinity of the peptide for its complimentary TCR
- Whether the peptide is similar to a self-peptide, whereby the parent of the complimentary TCR had already been removed via negative selection. Alternatively, inhibition by the peripheral tolerance CD4<sup>+</sup>CD25<sup>+</sup> cells (Shevach, 2002).
- A very unusual case where the peptide binds so strongly, that the CTL is sensitised for destruction by other CTLs; a process known as CTL fratricidal killing.

From the results of Ochoa-Garay *et al.* (1997), it can be deduced that a combination of MHC binding affinity and immunogenicity would be crucial in vaccine design. If the peptide is highly immunogenic, but exists sparsely on the cell surface due to low presentation yield, it would lower the efficacy of the vaccine. Conversely, if a peptide binds with a high affinity, but has low immunogenicity, the vaccine would also have lower efficacy. High affinity peptides might occupy a lot of "MHC space" on the cell surface and cause the other epitopes to fade into the background and escape immune surveillance. POPI is currently the only tool known to the author with reasonable immunogenicity prediction.

### Epitope Cross-Reactivity

Cytotoxic T-lymphocyte epitopes are usually polyspecific, meaning they stimulate the response of different T-Cell clones (Mason, 1998). Knowing there is only a limited set of CTL clones within the human body, an inference can be made that certain T-Cell epitopes are cross-reactive. Cross-reactivity, in the context of Cell-Mediated Immunity, is the definition given to an epitope that stimulates the response similar to T-Cell clones by a different epitope. In a recent study of cross-reactivity, it was determined that particularly the central part of an epitope, e.g. the threonine residue in SLYNTVATL is particularly sensitive to mutations abrogating cross-reactivity (Frankild *et al.*, 2008). As a consequence, the researchers developed a rudimentary measure to compare two epitope sequences, dubbed by this author as "The Frankild Score". However, cross-reactivity can sometimes be found between epitopes that share very little amino acids. For example, the epitope from Influenza A, GILGFVFTL has cross-reactivity with the Epstein-Barr

Virus epitope, SVRDRLARL. The Frankild score for these two epitopes is low, indicating that other subtle forces govern cross-reactivity that cannot necessarily be simplistically extrapolated from sequence similarity alone.

## 1.2 Tools that Predict Steps in the Antigen Presentation Pathway

In this section, three separate sets of tools will be discussed that cover each of the steps mentioned in antigen processing and presentation. They follow:

- Proteasomal cleavage - NetChop C2.0 and ProteaSMM (Kesmir *et al.*, 2002, Tenzer *et al.*, 2005)
- TAP affinity - The method by Peters *et al.* (2003)
- MHC binding affinity prediction by Bimas and NetMHC (Nielsen *et al.*, 2004a, Parker *et al.*, 1994)

### 1.2.1 Performance Measurements

Before discussing how these individual tools work, it is prudent to discuss how they are evaluated. The tools are usually evaluated in how they solve a binary problem, e.g. how proteasomal cleavage site predictor can distinguish between cleavage and non-cleavage positions. The following statistical measurements are very useful in determining performance:

1. True Positives (TP) - Predicted positives that are really positive
2. False Positives (FP) - Predicted positives that are really negative
3. True Negatives (TN) - Predicted negatives that are really negative
4. False Negatives (FN) - Predicted negatives that are really positive

From this, the following can be calculated:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (1.1)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (1.2)$$

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (1.3)$$

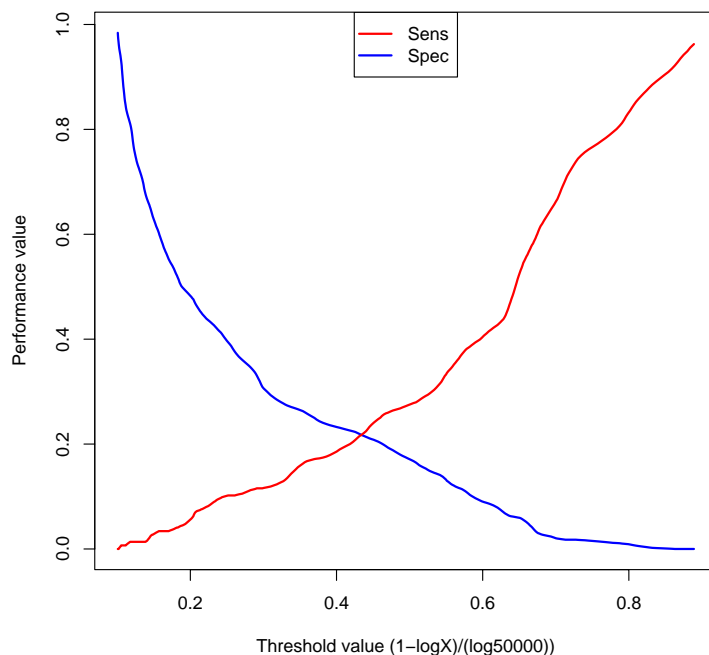


Figure 1.4: As the threshold values increase, it becomes clear that less positive value are predicted (drop in sensitivity) while the accuracy of the prediction (really the fraction of negatives predicted, specificity) increases.

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{FP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \quad (1.4)$$

Where sensitivity (Equation 1.1) and specificity (Equation 1.2) is the total fraction of TP and TN calculated, respectively. PPV (Equation 1.3) is the Positive predictive value and measures the fraction of correctly predicted positives. MCC (Equation 1.4) is the Matthews Correlation Coefficient (Matthews, 1975). The MCC ranges from a value of -1 to 1, 1 being a perfect prediction and -1 being an inverted prediction and 0 for a random prediction.

The values of these functions can be measured at different thresholds. This allows for the creation of a generalised performance measuring curve, called a Receiver Operating Characteristic Curve (Martin *et al.*, 2005). This curve is a plot of sensitivity *vs.* specificity (or, alternatively the False Positive rate,  $\text{FPR} = 1 - \text{Specificity}$ ). The area under this curve determines the overall performance of the prediction tool. The area is somewhere between zero and one, one being better performance than zero. An area of 0.5 constitutes a random predictor. Interestingly, if the tool produces a curve of an area below 0.5, say 0.3, it is merely an indication that the classification labels are reversed. For example, the tool reverses the predictions for a proteasomal cleavage site. Reversing the prediction calling would give this tool an AUC of  $\text{AUC} = 1 - 0.3 = 0.7$ . In Figure 1.4 we see the sensitivity and specificity values of an MHC binder predictor at different thresholds. The tradeoff between sensitivity and specificity is clear. In Figure 1.5 the ROC plots of different MHC binder predictors are superimposed, but instead of specificity, the aforementioned FPR is

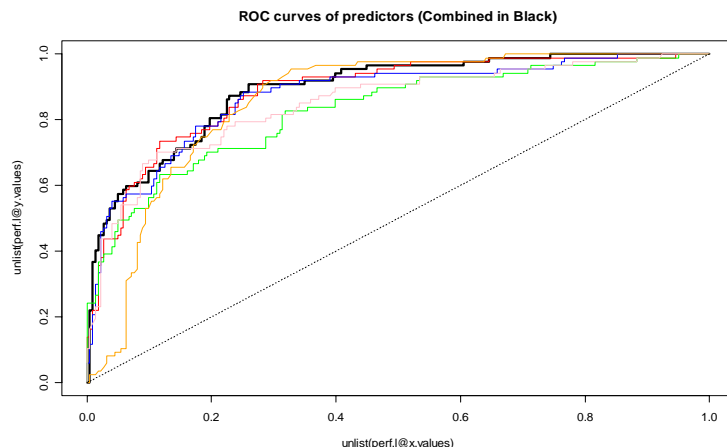


Figure 1.5: Curve depicting a ROC analysis of Sensitivity vs. False Positive Rate ( $1 - \text{Specificity}$ ). The different curves represent different tools used for predicting MHC ligands, with the solid black curve the result of a tool combining 5 predictors.

used, which still yield the same AUC values. The AUC also gives the average specificity and sensitivity values across all sensitivity and specificity values respectively. One might wonder what the big fuss is between a tool with an AUC value of 0.98 versus a tool with an AUC value of 0.99. Taking the average specificity as an example, there is only a 1.01 factor increase. However, specificity can also be used to calculate the FPR, which is a fraction loosely translated as "Given 100 positive predictions, what fraction of those should actually be negative?". Translating the specificity values to FPR values, the difference becomes more clear. The first predictor has an FPR of 0.02, while the second has an FPR of 0.01. This draws to the conclusion that the first tool makes twice the amount of false positive predictions than the second tool.

Different tools for the same prediction make use of training sets using different data scales and the AUC value really is a good way to compare them. Other ways to test the performance of a tool would be correlation and regression to measure how well the predictor's score correlates with a testing set.

### 1.2.2 Proteasome Cleavage Site Predictors

Two predictors will be discussed, namely NetChop C2.0 and ProteaSMM (Tenzer *et al.*, 2005). Both make similar assumptions, though constructed from vastly different data and methods. NetChop C2.0 uses a neural network approach, while ProteaSMM uses a scoring matrix. Both these methods investigate prior to training the importance of the residues at a cleavage site, plus the influence of surrounding residues. NetChop C2.0, however, made use of some clever training data. Very little data exists (or existed at the time of the tools' construction) pertaining to cleavage preferences for the proteasome. This is further complicated by having to predict cleavage sites for both the constitutive- and immunoproteasome. NetChop C2.0 takes advantage

Cleavage point  
▼

|                                      |   | P2 |    |    | P1 |    |    | P1' |     |     | B1A1' |     | Vw   |      |
|--------------------------------------|---|----|----|----|----|----|----|-----|-----|-----|-------|-----|------|------|
|                                      | O | A2 | B2 | C2 | A1 | B1 | C1 | A1' | B1' | C1' |       | O   | )    |      |
| ABC                                  | 1 | 1  | 0  | 0  | 0  | 1  | 0  | 0   | 0   | 1   | 0     | A2  |      | 10.0 |
| ABB                                  | 1 | 1  | 0  | 0  | 0  | 1  | 0  | 0   | 1   | 0   | 0     | B2  |      | 4.0  |
| BBB                                  | 1 | 0  | 1  | 0  | 0  | 1  | 0  | 0   | 1   | 0   | 0     | C2  |      | 6.0  |
| BCB                                  | 1 | 0  | 1  | 0  | 0  | 0  | 1  | 0   | 1   | 0   | 0     | A1  |      | -2.0 |
| <span style="color: red;">BBA</span> | 1 | 0  | 1  | 0  | 0  | 1  | 0  | 1   | 0   | 0   | 1     | B1  |      | 1.0  |
| CCA                                  | 1 | 0  | 0  | 1  | 0  | 0  | 1  | 1   | 0   | 0   | 0     | C1  |      | 2.0  |
| AAA                                  | 1 | 1  | 0  | 0  | 1  | 0  | 0  | 1   | 0   | 0   | 0     | A1' |      | 4.0  |
|                                      |   |    |    |    |    |    |    |     |     |     |       | B1' |      | 2.0  |
|                                      |   |    |    |    |    |    |    |     |     |     |       | C1' |      | 5.0  |
|                                      |   |    |    |    |    |    |    |     |     |     |       |     | -6.0 |      |

Taking the red dotted BBA as an example:

$$\begin{aligned}
 S(\text{BBA}) &= O + A2(0) + B2(0) + C2(1) + A1(0) + B1(1) + C1(0) + A1'(1) + B2'(0) + C2'(0) + B1A1'(1) \\
 &= 10.0 + 0 + 0 - 2.0 + 0 + 0 + 2.0 + 0 + 2.0 + 0 + 0 + 0 + 5.0 \\
 &= 17.0
 \end{aligned}$$

Figure 1.6: Example of an SMM calculation

of MHC ligands determined from different proteins. As whole proteins were used to produce the ligands, the flanking regions of the MHC binders could also be extracted from the sequence data of their original proteins. As mentioned in Section 1.1.1, little or no C-terminal trimming occurs, therefore the MHC binder was cleaved from its parent peptide at the C-terminal end. The authors also assumed that within MHC binder peptides have a very low probability of being cleaved, as they already escaped proteasomal cleavage.

## ProteaSMM

ProteaSMM is a matrix-based tool for predicting potential proteasomal cleavage sites (Tenzer *et al.*, 2005). It takes advantage of the Stabilised Matrix Method which is a type of position specific scoring matrix. Where it differs from the usual matrix-type calculations, is that it tries to reproduce experimental values from training data, and thus gives a quantitative result over a qualitative one. This is especially important in proteasomal cleavage predictions, since whether a protein will be cut at a specific position is more relative than it is absolute, as can be seen in its training data (Toes *et al.*, 2001). The data set contains not only the positions of cleavage sites, but also the frequency of peptides produced. From this, the prediction at a certain motif can be defined in a quantitative way. An example is displayed in Figure 1.6.

The full description of the SMM method can be found in the article by Peters and Sette (2005). Here, a brief overview of how the method was applied to predicting proteasomal cleavage sites will be explained.

- A matrix is constructed from sequence data using a window of length seven, designating  $P_n$  as residues before the cleavage point and  $P_n'$  as residues after.  $P_1$  would be the C-terminal residue of the cleavage point.
- For each training sequence, a binary value is assigned per position for the presence or absence of a residue
- A set of "paired residues" could also be defined, e.g an Ala at  $P_1'$  that co-pairs with an Arg at  $P_2$ . The algorithm checks whether there are enough incidences of a given pair, before it is actually included in the training
- The value of the amino acid is assigned a corresponding value from a weight matrix
- The weights  $V_w$  are adjusted using cross-validation, each time the algorithm attempts to minimise the error by adjusting the weights.
- When training is complete, together with the offset value, a prediction gives an estimate of the expected experimental value that could be obtained. The score obtained in ProteaSMM is a log-estimate of the total amount of fragments generated from a particular cleavage site.

ProteaSMM was trained on both constitutive proteasome data and immunoproteasome data. The AUC values for both sets were similar, ranging from 0.67 - 0.82 on different testing sets. It was noted in the article that it did outperform NetChop C2.0 and NetChop 20S which are explained in the next section.

### NetChop 20S and NetChop C2.0

At the time NetChop was designed, very little *in vitro* digestion data of the proteasome was available (Nielsen *et al.*, 2005). This is a huge problem, because in order for an artificial network to generalise a problem, it needs a large amount of data to perform with reasonable accuracy. Given enough data, an ANN's power comes from the fact that it can solve non-linear problems, allow for erroneous data and have the ability to "improve" itself (Brusic *et al.*, 2004). In the case of NetChop 20S and NetChop C2.0 (hereafter collectively referred to as NetChop), the researchers needed to increase the amount of available data. With the help of an CTL epitope database from whole, known proteins, they figured that the epitopes and the regions surrounding them could provide additional cleavage data, since a peptide will not be presented, unless it was cleaved at the correct place on the C-terminal end. However, as mentioned in Section 1.1.1, it is more likely for the immunoproteasome to be involved in fragment generation for MHC presentation, thus the additional data provided by the MHC ligands are more than likely biased towards the immunoproteasome cleavage preference and not the constitutive immunoproteasome. The

Table 1.1: . ProteaSMM exclusive residues in bold, NetChop exclusive residues underlined

| Position   | Positive Effect on Cleavage | Negative Effect on Cleavage |
|------------|-----------------------------|-----------------------------|
| <b>P1</b>  | <b>DFLY</b>                 | <u>PKGTN</u>                |
| <b>P2</b>  | <u>QYV</u>                  | <u>PD</u>                   |
| <b>P3</b>  | <u>V</u>                    | <u>GQ</u>                   |
| <b>P4</b>  | <u>PT</u>                   | <u>DK</u>                   |
| <b>P2'</b> | <b>DLH</b>                  | <u>KSREP</u>                |

difference in preference of these two proteasomes became clear in the Kullback and Lieber analysis of the cleavage sites. Briefly, the Kullback-Lieber distance is a log-odds ratio between particular residue at a particular position in the cleavage sequence and the same residue in the background. The researchers do say that this data should be approached with caution, since the MHC ligands mainly represent cleavage data for the immunoproteasome, not uncleaved sequences. Still, this tallies well with the earlier discussion that the constitutive proteasome prefers acidic residues at the C-terminal end of cleavage. Reading further, we are once again faced with a bit of a conundrum. According to the weights in the neural network after training on *in vitro* cleavage data on Enolase and  $\beta$ -Casein, the preferred residues at P1 (i.e. at the C-terminal end of the would-be fragment) are Phe, Tyr and Leu and disfavours Pro, Gly, Thr, Asn and Lys. Still, this is somewhat in accordance with the Kullback-Lieber distance results, but not really with the ProteaSMM constitutive matrix, where Asp is also a preferred amino acid at the P1 position. This is in accordance with theory. See Table 1.1 for a full difference between positional amino acid preference between ProteaSMM and NetChop.

## Neural Networks

Fully describing a neural network would take a full review on its own, but the gist of its operation is easy enough to explain briefly. Neural networks are designed to process information in parallel. This means, unlike a matrix-based method, a neural network would consider all the residues in all the positions of an imaginary sequence TCGGALL. (unless explicitly specified as is the case with SMM). This is achieved by assigning an input neuron to each position for each possible amino acid, so in this example it is  $7 \times 20 = 140$  neurons. Only the neurons corresponding to the amino acids at a position are activated. This 'signal' is passed to a hidden layer of neurons where weight adjustment occurs. The weights of the neurons are adjusted to try and minimise error with the desired output, also known as "back propagation". A significant parameter in a neural network is the learning rate. The smaller the learning rate, the smaller the adjustments are to the weights of the neurons and the converse is true for a large learning rate. The problem with a large learning rate is that the error between the predicted and desired result is not minimised. The problem with a small learning rate, is that the network takes too long to train. Usually, a differential learning rate is applied to compromise between the two factors. In immunological

tools, the output of the neural network is usually an approximation of the actual experimental result (Nielsen *et al.*, 2004a).

As is often observed when comparing prediction tools, the performance measure of a particular tool by different researchers are usually comparable. It is very tempting to come to the conclusion that the authors are biased, but hopefully this can be avoided. Performance of a predictor can be influenced by the set it was tested on. According to some (Tenzer *et al.*, 2005), the predictive performance in terms of AUC of NetChop averages 0.71, peaking at 0.78, while ProteaSMM averages at 0.72, peaking at 0.81. The authors of NetChop claim it to be 0.81. It does, however, make sense for ProteaSMM to perform better, since SMMs outperform neural networks on smaller data sets (Peters and Sette, 2005). The later version of NetChop, NetChop C3.0, (Kesmir *et al.*, 2002) does have a significant improvement over NetChop C2.0 with an AUC value of 0.86. The reason for the improvement has to do with generalising the input data for the network more, by using a BLOSUM50 matrix encoding of the amino acids (Henikoff and Henikoff, 1992), inclusion of HMM predicted inputs. Another significant difference between NetChop C2.0 and NetChop C3.0 is the inclusion of more proteasomal cleavage data. It is for this reason that NetChop C3.0 is also not compared to ProteaSMM.

One problem with using a ROC curve for analysing potential cleavage sites, is that a ROC analysis is of a binary nature. Subtle quantitative differences between the prediction tools are not taken into account. Furthermore, it would be unwise to predefine a threshold for defining a cleavage site, since it has been mentioned cleavage sites are not absolute. If there exists two consecutive cleavage sites, chances are that not one, but three possible products (all in different ratios) exist for the particular sequence (Peters *et al.*, 2002). With increasing amount of available data, prediction results should improve.

### 1.2.3 TAP Affinity Prediction

The affinity of a peptide to TAP can be a dramatic factor in delivering a potential MHC ligand to the MHC complex. There are various examples of MHC ligands that never become CTL epitopes (in this case, irrespective of immunogenicity) (Rammensee *et al.*, 1999). The reasons range from never being produced by the proteasome to poor delivery to the ER. As the proteasomal prediction has already been covered, this section deals with predicting the affinity of a peptide to the TAP complex. TAP affinity is strongly correlated with transport across the ER membrane into the ER lumen (Peters *et al.*, 2003).

**TAPSMM** The name assigned to the predictor is inventive, since no name was given for in the relevant article (Peters *et al.*, 2003). As can be deduced from TAPSMM, the method also relies on a Stabilised Math Matrix encountered in Section. It also uses three other matrices



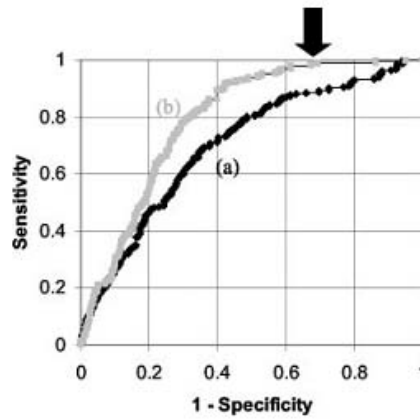


Figure 1.7: Curve (a) shows the ROC curve if N-terminal residues are not weighted, while curve (b) shows the result of changing the weighing parameter,  $\alpha$  to 0.2. (Peters *et al.*, 2003)

and the consensus matrix from them are used in predicting results. The researchers constructed their own TAP affinity dataset experimentally using IC50 values as the measurement. The set contained 430 9-mer peptides and 67 longer peptides. Using the 9-mer peptides, the matrices (of which two were of literature origin) were constructed from the amino acids at respective positions and the associated IC50 values. Calculating a score from it is a matter of summing the matrix values for each amino acid at each position. Binding of a peptide to TAP is mainly determined by the first three N-terminal residues and the C-terminal residue. Using this knowledge, the researchers could extend predictions to peptides longer than 9 amino acids, by using Equation 1.5:

$$t = \text{mat}_{1,N1} + \text{mat}_{2,N2} + \text{mat}_{3,N3} + \text{mat}_{9,C} \quad (1.5)$$

The results of the ROC analysis revealed that the method was fairly good, with an AUC value of 0.702. The researchers realise that the contribution of the N-terminal residues became less relevant with an increase in peptide length (i.e. window size). To compensate for this, a down-weighting factor was included for the N-terminal terms in Equation 1.5, leading to Equation 1.6:

$$\bar{t} = \text{mat}_{9,C} + \frac{\alpha}{L-8} \sum_{k=1}^4 (\text{mat}_{1,N1} \text{mat}_{2,N2} + \text{mat}_{3,N3}) \quad (1.6)$$

It was determined that the optimal value for  $\alpha$  was 0.2. This increased the AUC value of 0.702 to 0.792. The difference in the ROC curves are shown in Figure 1.7.

Interestingly enough, there is some correlation between the value of  $\alpha$  and the MHC allotype to which a peptide will bind. See Table 1.2.

Table 1.2: It is interesting to note how the value of  $\alpha$  and consequently the N-terminal has an effect on binding to specific allotypes of HLA (Peters *et al.*, 2003).

| Allele    | Optimal $\alpha$ for a window size of 10 |
|-----------|--|
| HLA-B44   | 0  |
| HLA-A24   | 1  |
| HLA-A3    | 1.2                                      |
| HLA-B27   | 4  |
| HLA-A0201 | 0.4                                      |

### 1.2.4 MHC Ligand Prediction

Various tools exist nowadays for predicting MHC ligands. In the antigen presentation pathway, this is the most selective step for various reasons and to an extent, the most problematic. MHC molecules are quite selective in what peptides they bind and with what affinity they bind (Rammensee *et al.*, 1995). This is determined by the residues forming the groove, which form certain binding pockets. There are pockets for anchor residues, which are essential for a ligand to bind to the groove. However, these anchor residues and positions along the groove where they occur vary between the MHC types (e.g. HLA-A, HLA-B, HLA-C) and allotypes (e.g. HLA\*A0201, HLA\*A0204). Because of the large variability, there needs to be huge amounts of data available in order to find a reasonably accurate prediction method.

As the amount of binder data became more and more available, so did the MHC ligand predictors (in general) improve. From the early Bimas method (Parker *et al.*, 1994) to the present day methods like the Support Vector Machine (SVM) method, SVMHC (Dönnes and Elofsson, 2002, Dönnes and Kohlbacher, 2006), the ANN based method of NetMHC (Nielsen *et al.*, 2004a) and the SMM based methods (Nielsen *et al.*, 2004b) and others such as SYFPEITHI (Rammensee *et al.*, 1999) and PepVac (Reche and Reinherz, 2005). Bimas and NetMHC will be discussed as they are based on different techniques and assumptions. Bimas uses a method that treats each position in the ligand as independent of all the others, while NetMHC reveals to us that there is, in fact, a lot of interdependence between the residues of a ligand. It is by this assumption that NetMHC performs a lot better than Bimas.

#### Bimas

Bimas was one of the original tools for predicting MHC ligands. It is based on the assumption that binding of a peptide to MHC is independent with respect to the residues it is made up of and that the total binding can be expressed as an additive function. The binding affinity is correlated with the half-life of the binding of  $\beta$ 2-M to the MHC complex. The peptides used for data set construction were carefully constructed with substitutions occurring at strategic positions and sometimes, by looking at the set closely, substitutions are progressive. At single substitutions,

the half-life was compared and a coefficient was assigned, e.g.

$$C_{A_7, E_7} = \frac{T_{GLFGGAGV}}{T_{GLFGGAGV}} = \frac{770}{530} = 1.45$$

There are a total of  $20 \times 9 = 180$  variables (coefficients) to consider, which, for the authors, was a lot given the amount of data they had. It was decided that if a particular positional amino acid's coefficient was too close to 1.0, it would remain fixed at 1. Also, if there was no way to measure the coefficient of a variable at a certain position, it was fixed at 1.0. A total of 82 variable coefficients were allowed to be changed during optimisation. The output of the function is an approximation of the experimental  $\beta$ 2-M half-life data. The prediction seems to correlate reasonably well with the testing half-life values.

The authors note, however, the limitation of the independent binding assumption, by listing a few peptides with curious properties. Still, evaluating this method on current data sets yields pretty impressive results, e.g. when tested on the HLA\*A01 (Tenzer *et al.*, 2005) and HLA\*A0201 (Trost *et al.*, 2007), AUC values of 0.9934 and 0.920 are obtained, though the accuracy of performance analysis of HLA\*A01 seems a little doubtful<sup>1</sup>. We will see in NetMHC how the shortcomings of Bimas (and other tools) were addressed.

## NetMHC

During the following years after Bimas, a flood of MHC data became available. It was the start of a real feast for people eager to solve the MHC ligand problem. The databases grew, more and more became known of the specific binding motifs for various MHC allotypes. The problem was, however, that the data was of a binary nature and not the sleek quantitative data used in construction of Bimas. Even so, discrete datasets like SYFPEITHI (Rammensee *et al.*, 1999) and MHCPEP (Brusic *et al.*, 1998) still provide very useful information. MHCPEP also includes immunogenicity data of MHC ligands. Other data sets were constructed using quantitative IC50 values. The IC50 value in terms of MHC binding is a measurement of what concentration peptide there needs to be for a specific amount of MHC molecules to be 50 percent saturated (Sidney *et al.*, 2001, Sylvester-Hvid *et al.*, 2002). The reason for the necessity of a quantitative data set is to distinguish between "Is it a binder?" and "Precisely how much of a binder is it?". This could aid especially a learning method such as ANNs to add weights to the neurons, because even if two peptides do bind well to MHC, one of them is likely to be "more or less of a binder".

NetMHC makes use of the Sette and SYFPEITHI data sets in training of the neural network. The authors elegantly demonstrated why they think an ANN would be more appropriate in solving the MHC ligand problem. The theory is that interdependence between residues in a peptide could have a dramatic effect on MHC binding. By taking MHC ligands from the database, they constructed a  $9 \times 9$  matrix with each position represented by a score obtained from using

<sup>1</sup>Why would better tools be constructed if Bimas already performed at a near perfect level?

Equation 1.7.

$$M_{ij} = \sum P_{ij}(\mathbf{ab}) \log\left(\frac{P_{ij}(\mathbf{ab})}{P_i(\mathbf{a})P_j(\mathbf{b})}\right) \quad (1.7)$$

Where  $i$  and  $j$  are positions of which a mutual information score is calculated, with  $\mathbf{a}$  and  $\mathbf{b}$  representing amino acids at  $i$  and  $j$  respectively.  $P_{ij}(\mathbf{ab})$  is the probability of simultaneously having amino acid  $\mathbf{a}$  at  $i$ , while having  $\mathbf{b}$  at  $j$ .  $P_i(\mathbf{a})$  represents the total occurrence of  $\mathbf{a}$  in the background (same applies for  $P_j(\mathbf{b})$ ). NetMHC makes use of two ANNs for prediction, by summing the result of the two in a weighted fashion. There are ten input neurons. Nine for the positions (when training 9-mers) and the tenth for the output of a HMM. One ANN is trained using sparsely encoded amino acids, which means that only one amino acid neuron is activated per position at a time and the BLOSUM neurons can be activated based on the similarity between the amino acids at a given position. The limit amount of data makes it wise to rather encode the sequences using a BLOSUM50 encoding matrix. The BLOSUM matrix is a matrix obtained from observing substitution frequencies of amino acids from aligned sequences in a LOD fashion (Henikoff and Henikoff, 1992). This reveals subtle relationships between the amino acids. An example is given in 1.8. Although there are two mismatches between sequences RFFIVDKLL and RFFLVEKLL; the network, having already seen RFFIVDKLL, assumes the P3 amino acid is the same. The sparse encoding sees them as different and has to adjust the weights of two variables. So, why not only use the BLOSUM encoding? The BLOSUM encoding will cause the network to behave a little more erroneously in comparison with the sparse encoding, when faced with the sequences RFFIVDKLL and RFFLVDKLL. Here, only the P3 residues are different, but the BLOSUM encoding will once again see it as the same. Therefore, the weighted output yields a significantly better result than either of the encoding methods alone. Inclusion of the HMM output also proved fruitful.

Although this method is powerful and tested to be the most accurate with an AUC for HLA\*A0201 of 0.932, ANN training requires a lot of data. Here, by also including the SYFPEI-THI data, the researchers were able to increase the set. The amount of quantitative data for the other alleles in comparison with HLA\*A0201 is a lot less and prediction performance should accordingly be lower.

### 1.3 Modeling the Entire Process

There have been attempts to model the Class I Restricted pathway (Tenzer *et al.*, 2005, Larsen *et al.*, 2005). Of the three classes of tools, namely MHC ligand predictors, TAP affinity predictors and proteasomal cleavage site, MHC ligand prediction is the most accurate. This is true for a number of alleles. There have even been attempts to design tools that predict epitopes covering

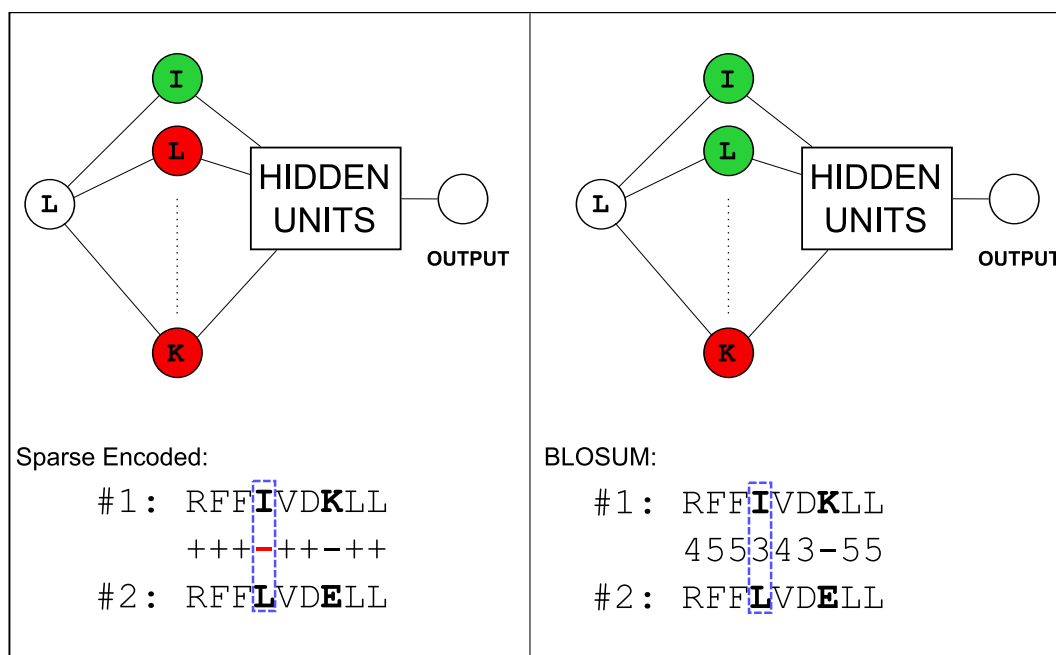


Figure 1.8: This Figure depicts how BLOSUM encoding could help the ANN to generalise a problem. The two peptides have two mismatches between them at P4 and P7. With sparse encoding, the ANN now has to take two mismatches into account for weight balancing. With BLOSUM encoding, the Ile and Leu of P3 are seen as the "same" amino acid and the network only has to adjust for the one P7 value as it has "seen" a Leu before (i.e. the Ile)

95% of the American population (Reche and Reinherz, 2005). Prediction of epitopes for a specific allele seem fruitless and some prediction tools, like the newer versions of NetMHC allow for supertype prediction, i.e. across the entire HLA-A2 serotype. Still, further binding data on other alleles, especially neglected population groups should be gathered. The current MHC ligand predictors are also capable of predicting ligands in the size ranges of 8-11 amino acids. It seems surprising that there aren't a lot of data on *in vitro* digests of the two classes of proteasomes. Within species variability of these is negligible and good methods have already been established to determine cleavage sites. TAP prediction is fairly accurate and proven to increase the accuracy of MHC ligand binders by filtering TAP non-binding sites and thereby eliminating the MHC ligand that goes with it.

It is the author's opinion that predictors should not be binary classifiers. Quantitative values at each step could be invaluable when approximating a candidate CTL epitope for a vaccine. A peptide sequence that has a positive result for all the tools may not necessarily be a good CTL epitope (excluding the factor of immunogenicity here); If 10 fragments of the same sequence containing the CTL epitope are produced, but only 0.7 of that transported across the ER by TAP and only 0.5 of that presented to the MHC molecule, that means only 4 of the original 10 potential epitopes were actually presented.

A lot of work need be done in improving the prediction performance of especially the proteasomal cleavage site predictors. Combining the results of predictions at different steps in the

antigen presentation pathway is attractive, but not when the error accumulates at each step, bringing it again ever closer to a random prediction. One of the purposes of these prediction tools would be to decrease experimental costs. By screening only 10 epitopes for antigenicity instead of hundreds, the throughput of the experimental method is increased ten fold. For this to happen, though, the predictors need to be of high quality.

The troublesome aspect of training MHC predictors is the nature of the dataset. Using a simple PSSM based on LOD scores of binders and non-binders, the author designed a tool with an AUC value of 0.88. This was surprisingly high and even more surprising was the fact that similar AUC values could be obtained when at a training- and testing set size ratio of 2:8. It became clear, however, that the nature of the dataset was to blame for the unrealistic performance. When testing the same method on a different set, AUC values of only 0.72 could be obtained. Caution should therefore be taken when choosing training and testing sets.

### 1.3.1 Other Available Tools

In conjunction with resources like SYFPEITHI and MHCPEP, other online resources are available, such as The Jenner Institute (<http://www.jenner.ac.uk>), the International Immunogenetics Information Management System (<http://imgt.cines.fr>) (Lefranc, 2005) and the online community resource for computational immunology (Peters *et al.*, 2006), EpiMHC <http://immunax.dfci.harvard.edu/bioinformatics/epimhc/> (Reche *et al.*, 2005), MHCBN <http://www.imtech.res.in/raghava/mhcbn/> (Bhasin *et al.*, 2003).

## 1.4 The other Immunological Responses

Investigating tools for the other classes of immune responses is beyond the scope of this review. It is worthwhile to mention that vast improvements have been made in predicting MHC Class II ligands. Helper T-Cells (CD4+, HTL) need to be activated for the CTL to differentiate into memory cells (Shedlock and Shen, 2003). As for the humoral immunity, predicting B-Cell epitopes, i.e. areas on a molecule where an antibody can bind is an extremely difficult task. Even with the like of BepiPred which predict continuous epitopes (epitopes formed by a continuous range of amino acids), prediction accuracy is very close to random (Larsen *et al.*, 2006). Hopefully, as more structural data becomes available and better structural prediction algorithms, the performance of BepiPred could increase.

## 1.5 Problem Statement

Phylogenetic methods are currently used to assess the genetic distance between strains of pathogenic organisms, and by extension, immunological distance. In the context of Cell-Mediated Immunity, specific differences in epitope repertoire between the pathogen strains may be an improved way to determine immunological distance. To the author's knowledge, there currently exist no available and freely accessible methods to perform this task. Using quantitative prediction methods for proteasomal cleavage, TAP affinity, MHC affinity, Immunogenicity and cross-reactivity, a system has been created to assess the immunological differences between strains of pathogens. The result of which is Fortuna, a freely accessible web-based tool to perform these immunological analyses.

## 1.6 Aims

1. Combine the results from various computational immunology tools concerning the Class I restricted antigen presentation pathway
2. Develop a novel TAP-ligand predictor
3. Devise a way to compare different sets of epitopes by using a combination of results from the antigen pathway prediction tools and the Frankild score
4. Design a web-based interface to access these tools
5. Perform a small study on the CTL epitope profiles of the Human Immunodeficiency Virus and Influenza A

## Development of Fortuna

Fortuna is a tool designed to aid the meta-analysis of immunological properties between variants of the same protein sequence. To achieve this, it utilises a combination of prediction tools for each of the steps in the Class-I restricted antigen presentation pathway (hereafter C1APP) as well a method to approximate cross-reactivity. Additional analysis on epitopes include frequency versus entropy analysis as well as self-epitope analysis. In this chapter, the design and implementation of tools used to facilitate the aforementioned will be described. In 2.1 the method by which C1APP predictors are implemented is shown. Proteasomal cleavage and MHC ligand affinity predictions are discussed in Sections 2.1.1 and 2.1.3. A novel TAP predictor, Variable Lengthed TAP Predictor (hereafter VLTAPP) has been developed and is discussed in Section 2.1.2. Beyond using VLTAPP for pathway prediction purposes, the author wishes to illustrate the advantages and challenges associated with immunological prediction tools. Thus, a complete description of design, training and validation of the predictor is provided. The method to combine the prediction scores is discussed in section 2.1.4. In contrast to other prediction methods, which in the whole simply add the individual predictions of proteasomal cleavage, TAP affinity and MHC affinity together, here a method is described that systematically produces a final score from the results of the individual predictors. Analysis and visualization procedures are discussed in section 2.3. This section will focus on clustering analysis of sequences in Section 2.3.4, analysis of epitope sequence entropy and frequency in Section 2.3.1 and self-epitope discovery. Visualization procedures are given in conjunction with the description of the analysis procedures. The implementation of Fortuna as a web-based application is described in the next Chapter and will illustrate the interface to the analysis and prediction tools. Overview of the development process is shown in Figure 2.1 on the following page.

### 2.1 Pathway Predictions

This section deals with the different predictors used in epitope prediction. First, the proteasomal cleavage methods will be discussed and how the results from this prediction is used to prepare



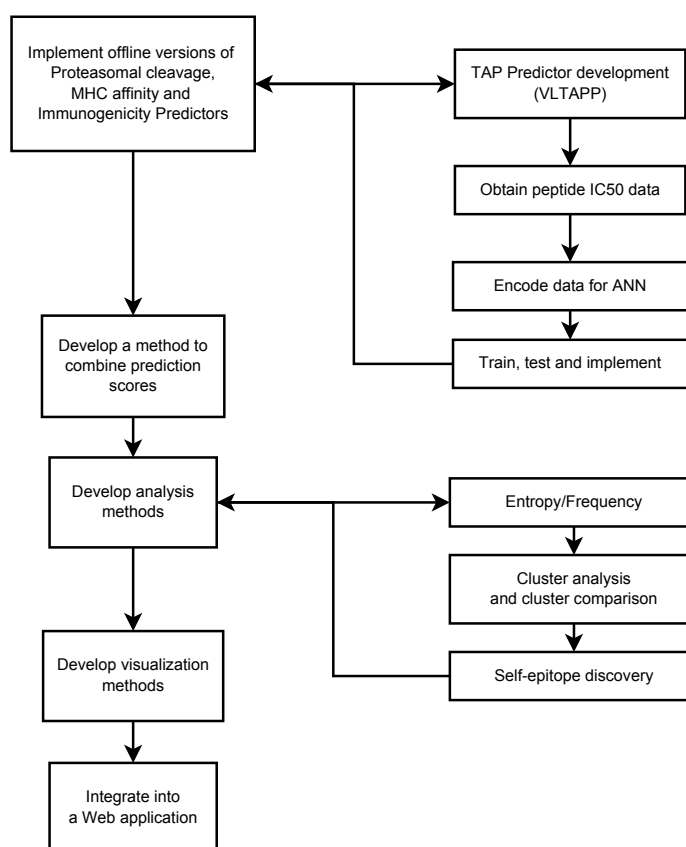


Figure 2.1: Overview of the development process of Fortuna. First, offline versions of NetMHC (MHC prediction), ProteaSMM (proteasomal cleavage prediction), and POPI (immunogenicity prediction) were implemented. The TAP predictor, VLTAPP was constructed from the obtained data and all prediction tools were integrated. Analysis tools and visualizations were developed as shown in the diagram and subsequently integrated into a web-based application, called *Fortuna*.

ligands for the next step, TAP ligand prediction. A novel predictor for TAP is constructed and the design as well as implementation of it is discussed. Lastly, the implementation of MHC affinity and peptide immunogenicity is discussed. Finally, a method is shown to combine the predictions to give a final epitope score.

### 2.1.1 Proteasomal Cleavage Prediction

Proteasomal cleavage is the very first step in the Class I-restricted antigen presentation pathway. It is also the first step in the simulated version of the pathway. There are many proteasomal prediction tools available, however, very few that can predict cleavage sites in a quantitative manner, i.e. how many times a certain site in a protein will be cleaved relative to other cleavage

sites. ProteaSMM method that does quantitative predictions on proteasomal cleavage sites (Tenzer *et al.*, 2005). It is based on PSSM that does predictions for both the constitutive and immunoproteasome. NetChop is another prediction tool that was considered (Nielsen *et al.*, 2005), but it differs from ProteaSMM by two very important factors:

1. Predictions are discrete
2. Does not distinguish between cleavage sites for the Immuno- and constitutive proteasome

The goal for the final output where Proteasomal cleavage, TAP affinity, MHC affinity and immunogenicity are all taken into account is to not only determine whether a potential MHC ligand will be displayed on the cell surface, but also the amount of it that will be displayed. It will be fruitless to consider an appropriate MHC ligand as an immunological target if its availability from the proteasomal cleavage step is low. Furthermore, there is a significant difference in the amount as well as composition of the proteasomal digests between the two main proteasomal types. Since TAP is dependent on the products of the proteasome, the difference in amount as well as actual fragments produced will be determined by the type of proteasome employed by the cell. NetChop does not reveal this discrepancy. The question remains, however, how to determine in a quantitative way the fragments passed to TAP.

### Quantitative Calculation of Proteasomal Fragments

Initially, proteasomal cleavage prediction is performed independently across all the positions in the input peptide. There is no information on the relative amount of a given sub-fragment between two cleavage positions. The score output of ProteaSMM can be interpreted as a relative amount of cleavage at a site and can thus be used to calculate the amount of a fragment existing between two sites. To achieve this, a probabilistic method is created. Assuming the maximum value possibly produced by an ProteaSMM matrix relates to a 100% probability of a site being cleaved, all the scores can be converted to a pseudo-probability. The procedure is:

1. Determine the ProteaSMM scores for a given sequence
2. Calculate the cleavage positions
3. Obtain the maximum possible value for the matrix used in prediction
4. Convert the scores to a fraction of the maximum score

With the cleavage probabilities calculated, fragment probabilities can now be determined. In simple terms, given the probability of cleavage sites  $A$  and  $B$ , what is the probability of fragment  $AB$ ? More formally described in Equation 2.1 on the next page

$$P(AB) = P(A) \times P(B) \quad (2.1)$$

Where  $P(AB)$  is the probability of a fragment formed by sites  $A$  and  $B$ ,  $P(A)$  and  $P(B)$  are the probabilities for a cleavage site to occur at position  $A$  and  $B$ . This assumes that the cleavage events  $A$  and  $B$  are independent, i.e. the existence of one cleavage site does not inherently influence the base probability of another cleavage site. The next step is to determine the probability of the fragment  $AB$  given another cleavage site exists between them. How much will the internal cleavage site,  $C$  influence the probability of  $AB$  being a product of proteasomal digestion? What is the probability that the cleavage of  $AB$  will occur while  $C$  does not? The procedure is relatively simple and based on the rule of conditional probability and shown in Equation 2.2.

$$P(AB|C') = P(A) \times P(B) \times P(C') \quad (2.2)$$

Where  $P(C')$  is equal to  $1 - P(C)$ . Given that  $B$  exists on the C-terminal end of the fragment  $AB$ , the probabilities of all the fragments formed by cleavage sites between  $A$  and  $B$  are calculated. There are two limits imposed on this procedure:

1. Minimum length of the fragment is 9 amino acids
2. The maximum length of the fragment is defined by the user, usually 20

The problem with the first condition is that a cleavage site with high probability may occur within a region defining an MHC ligand. The reason for disregarding this potential problem is because of the unknown nature of quantitative prediction performance of ProteaSMM; including more predicted cleavage sites may increase overall error of fragment probability. The maximum limit is imposed because the training set of the predictor does not contain peptides of length greater than 17 amino acids. An example of the procedure is given in 2.2 on page 27. Since proteasomal cleavage is also a rate-length dependent process, the probability of very long peptides will become negligible; more internal cleavage sites mean less probability for fragments formed by the terminal ends tested. TAP ligands also become less potent as their length increases. The exception to the maximum length rule is when there are no cleavage sites within the length limit upstream from a particular cleavage site. The solution is to find the next available upstream cleavage site. The rationale behind using the C-terminal cleavage site and working upstream instead of starting at the N-terminal cleavage site and working downstream, is that there is very little evidence of C-terminal cleavage beyond the proteasome (Snyder *et al.*, 1994). In the ER, where loading of a peptide onto MHC occurs, the ERAP molecule is designated to trim the fragment from TAP to appropriate size. Thus far, overwhelming evidence suggests that N-terminal proteolytic cleavage by the ERAP molecule is the *de facto* way of doing this.

Restrictions are therefore applied to the C-terminal end of a potential MHC ligand, while the N-terminal end is allowed to be more variable.

With careful investigation, it can be determined that the fragments probabilities calculated in 2.2 on the following page do not add up to the probability of the C-terminal end when summed as they should when consulting probability rules concerning independence. To compensate, the probabilities are multiplied by a factor as shown in Equation 2.3.

$$F_{NC} = \frac{\sum_{k=C_2}^N P(Ck)}{P(C)} \quad (2.3)$$

Where  $F_{NC}$  is the factor that all the probabilities of the fragments should be multiplied with,  $P(Ck)$  is the probability of a fragment defined by cleavage point  $C$  and the upstream cleavage point  $k$ ,  $C_2$  is the first cleavage point upstream from  $C$  and  $N$  is the most upstream cleavage point. An illustration of the entire process is shown in Figure 2.2. The author is aware of the proteasomal cleavage prediction tool, FragPredict (Holzhütter *et al.*, 1999). This tool does a two-run pass that consists of first predicting the cleavage sites and then the possibility of fragments therein. However, implementation of this method is exceptionally difficult, it does not distinguish between immuno- and constitutive proteasome cleavage sites, and subsequent methods also claim to perform better in terms of cleavage prediction (Nielsen *et al.*, 2005, Tenzer *et al.*, 2005, Ginodi *et al.*, 2008).

### Implementation of ProteaSMM

A local version of ProteaSMM is available. The predictions are made on 10-mer sequences with the cleavage point existing between the 6th and 7th amino acid. The input sequence(s) are scanned and a non-redundant list of 10-mers are built as shown in Figure 2.3 on the following page. The word list is passed to the SMM application provided. Various matrices trained on different data are provided for both prediction of constitutive- and proteasomal cleavage sites. The enhanced matrices were used, since they were trained on the most amount of data and are, theoretically, the most accurate. The output from SMM as a rule is a text file containing a list of words (peptides) with corresponding predictions values. For ProteaSMM the values are the  $\log_{10}$  quantitative values that equate to relative amounts.

#### 2.1.2 Variable Lengthed TAP Predictor

##### Construction of VLTAPP

There are a multitude of tools available concerning the prediction of ligand affinity to TAP. They include PREDTap (Zhang *et al.*, 2006), Tappred (Bhasin and Raghava, 2004) and the SMM Method by Peters (Peters *et al.*, 2003). The different methods employed are attuned

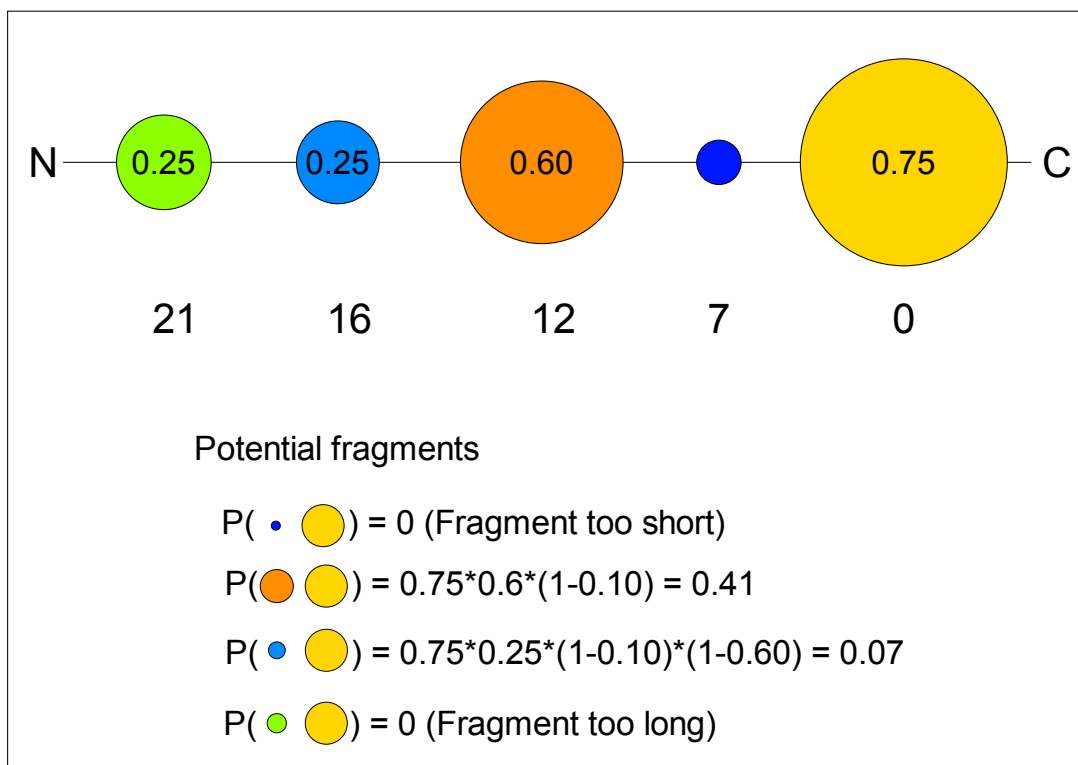


Figure 2.2: The Figure illustrates a region of a protein sequence containing predicted proteasomal cleavage sites. The size of the circles are representative of the cleavage probability. The letters *N* and *C* represent the N- and C-terminal ends respectively. The numbers below the circles indicate the position upstream from the C-terminal end. The fragments C-7 and C-21 are not viable since they fall beyond the range limit of allowed lengths, namely [9,20]. The two viable fragments, C-12 and C-16, have their probabilities depicted on the Figure. To illustrate the influence of length on the procedure, we can insert a hypothetical cleavage position, C-14 (not shown), with a probability equal to C-12. This would reduce the score of C-16 by a factor of 0.40, which would make it 0.03.

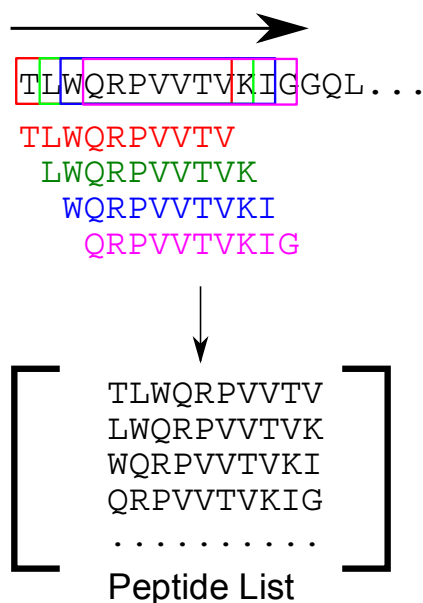


Figure 2.3: An input sequence is read from N- to C-terminal. Words of a predefined are extracted from the sequence and added to a list. Only words that have not been added to the list are considered to remove redundancy in subsequent predictions performed on the words.

Table 2.1: The size of the dataset containing TAP ligands of varying sizes with all possible single amino acid substitutions to a single reference peptide.

| Length of Peptide Ligand | Possible Substitutions | Total |
|--------------------------|------------------------|-------|
| 8                        | $8 \times 20 = 160$    | 160   |
| 9                        | $9 \times 20 = 180$    | 340   |
| 10                       | $10 \times 20 = 200$   | 540   |
| 11                       | $11 \times 20 = 220$   | 760   |

to the type of prediction made. Some focus on classification, i.e. making discrete predictions whether an input peptide would bind to TAP with an appreciable affinity, while others use regression methods to make quantitative predictions of the peptide's binding affinity to TAP. As was demonstrated with MHC ligand prediction, peptide lengths vary between 8-11 amino acids in the majority of cases and there is an adequate amount of information for many (but in no means all) HLA allotypes to make at least moderately accurate predictions across all lengths. However, since TAP ligands are the product of proteasomal digestion, the length of the peptides vary considerably. This poses a problem for anyone wishing to construct a TAP prediction tool, because the size of the training set increases quadratically with every additional lengthed peptide included in the set. It can be demonstrated by constructing a hypothetical a training set consisting of variable length peptides with each length containing all the possible single amino acid substitutions to a reference peptide. See Table 2.1.

Fortunately, only the terminal portions of the ligand are important in TAP binding. Most notably, as deduced in the literature, the three N-terminal amino acids and the C-terminal amino acid (van Endert *et al.*, 1995, Uebel *et al.*, 1997, Daniel *et al.*, 1998). The length of the ligand is inversely proportional to the influence of the N-terminal end amino acids on binding. The method by Peters makes use of a position specific scoring matrix in conjunction with a summation term for the three N-terminal amino acids. A factor is applied to the N-terminal scores to compensate for the ligand length (Peters *et al.*, 2003). This method separates the training of the 9-mer ligands and longer ligands. Here, length will directly be included in the training of the ANN to be constructed.

**Data Acquisition and Pre-processing** The data for the TAP predictor training set was obtained from the AntiJen web service (Toseland *et al.*, 2005). Ligand data criteria are variable and only those entries satisfying the following criteria were considered:

1. Measurement in IC50 values
2. Use of the standard peptide RRYNASTEL
3. Ligand for the human TAP

Table 2.2: Adjustment of IC50 values from entries using different standard peptide concentrations.

| Peptide   | Std Peptide Concentration | IC50 | Adj IC50 | log <sub>2</sub> IC50 |
|-----------|---------------------------|------|----------|-----------------------|
| AAASAAAAK | 250                       | 2143 | 2143     | 11.065                |
| AAASAAAAK | 150                       | 1709 | 2983     | 11.543                |
| AWASAAAAY | 250                       | 60   | 60       | 5.907                 |
| AWASAAAAY | 150                       | 45   | 75       | 6.229                 |

The IC<sub>50</sub> value represents the inverse affinity of TAP and is a close approximation to the dissociation constant  $K_d$  (Barlow *et al.*, 1997). The measurements are carried out by adding varying concentrations of a testing peptide to a solution and then measuring how much of a reference peptide, whose concentration remains constant, managed to bind to TAP. The results are plotted and through interpolation, the concentration of the testing peptide needed to block 50% of the available TAP molecules is determined. For example, 83 mM is needed for AAAAAAAAY to bind to 50 percent of the TAP molecules while a concentration of 250 mM RRYNASTEL exists in the solution (van Endert *et al.*, 1995). The amount of standard peptide used in different experiments can vary. This causes the IC<sub>50</sub> values to be seemingly highly variable. To compensate, the fold difference between concentrations of the standard peptide and testing peptide were measured, approximating the values to what they are if 250 mM RRYNASTEL were used. Examples of this procedure is demonstrated in Table 2.2. For one ligand that has multiple experimental values, the adjusted results were averaged. The final set consists of 343 peptides. This set contains less 9-mer peptides than used in other studies. After investigation, it was discovered that some entries in AntiJen did not meet the aforementioned criteria to be included in the set, nor were all the peptides from the single amino acid substitution set available.

**Input Data Encoding** For an ANN to perform with appreciable accuracy, the input nodes (parameters) of the network need to be appropriate. In the context of predictions made from amino acid sequence, each input node can be a binary value representing a unique amino acid at a given position. For instance, in a 4-mer window, 80 input nodes can be used to represent each of the 20 amino acids at a given position. The limited and missing data in the TAP training set makes the use of this binary definition of the amino acids problematic, e.g. if an amino acid at a particular position isn't included in the set, there is no way for the neural network know what its influence would be on the output. Therefore, it is prudent to use physiochemical properties to make an approximation of an 'unseen' amino acid. Properties used to encode an amino acid to a set of numerical values include, but are not limited to, hydrophobicity, pKa values for the side chains, volume and structural nature (cyclic/aliphatic). An example of this would be the PredTAP tool that uses a multitude of amino acid properties. A more indirect method is to make use of a BLOSUM matrix to encode the amino acid. This is done by making 20 value

Table 2.3: Amino acid properties used as input parameters

| Property               | Value  | Description  |
|------------------------|--------|--|
| Cyclic                 | {0,1}  | Has a Cyclic R-group, such as F,Y,W,H  |
| Aliphatic              | {0,1}  | Has an Aliphatic R-group, such as K,L,V,I  |
| Hydrophobicity         | {0,1}  | Utilises Kyte-Doolittle Method (Kyte and Doolittle, 1982)  |
| pKa Side Chain         | {0..1} | Range Scaled value of pKa values. For acids, the values [1..7] are reversed and scaled to [1..0] and Bases [7..14] are scaled to the values [0..1] |
| Volume                 | [0..1] | Volume of the R-group  |
| Proline                | {0,1}  | Discrete value to indicate amino acid is Proline   |
| Isoelectric point (pI) | {0,1}  | Ranged scaled value of the pI value for the amino acid   |

vector of BLOSUM scores for the amino acid versus itself and all the other amino acids in the matrix. This method was used in MHC ligand predictor, NetMHC, with great success. This encoding scheme was also investigated, however, the nature of each favoured binding residue will be of interest here (or comparison to other studies examining the nature of TAP ligands. See Table 2.3 a list of properties used.

Even though values such as pKa have a range limit of [1,7] for acids and (7,14] for acids, the values for each property in the encoding matrix fall in the range [0,1]. The reason for this is to assess the contribution of the physiochemical properties on the output score without scale bias. Input parameters that have a greater value may be seemingly down-weighted while those with a small value range may be up-weighted. Scaling solves this problem.

**Artificial Neural Network Construction and Training** The AMORE package allows for the construction of a simple feed-forward ANN in the R statistical language (Limas *et al.*, 2007). It is implemented in the *R programming language*. The input nodes of the neural network was defined by:

1. The aforementioned physiochemical properties of the four C-terminal and the four N-terminal residues
2. The length in amino acids of the peptide transformed to  $\log_2$ .
3. The average physiochemical properties across the whole peptide (only hydrophobicity, pI, Proline and pKa values used)

Central to the training of an ANN, a choice has to be made on the amount of hidden neurons, the learning rate and the number of training cycles. With each iteration, the weights of the nodes



are adjusted to better fit the desired output. A careful balance has to be achieved between the accuracy of the network on the training data and testing data. Too many training cycles will lead to over-fitting whereby the network might perfectly explain the training data, but fail to perform reasonable predictions on new data. On the other hand, if too little training cycles are performed, the network tends to under-fit the data leading to similar erroneous predictions. The same scenario is encountered with improper selection of hidden neuron functions and/or the amount of hidden neurons. Too many hidden neurons also lead to over-fitting. The learning rate is another point of interest. If the learning rate is too large, the network cannot optimize itself due to too large weight changes constantly 'missing' the optimal weights. If the learning rate is too small, the network will take eons to reach the optimal weights. For these two reasons, a momentum is added to the training cycles with the amount of weight change decreases per cycle, meaning that 'rough' adjustments are made to the weights in the earlier training cycles while finer tuning is done to the weights in the later cycles as the error rate decreases.

### Implementation of VLTAPP

Including TAP prediction in any simulated Class I restricted antigen presentation pathway is a unique problem. It is the only step that is guaranteed to depend on the output of the previous step, namely proteasomal prediction. Whereas prediction of proteasomal cleavage sites, MHC affinity and immunogenicity can be done independently, TAP prediction input is generated from proteasomal cleavage prediction output. In other research, ligands equal to MHC ligand length are assumed to be the definitive TAP ligand required for transport of the pro-MHC ligand to the ER where it gets loaded onto MHC. This assumption does not hold much water, because many lengthed proteasomal products can be produced. The extraction procedure is shown in Table 2.4 on the following page. A unique list of proteasomal products are encoded to the appropriate VLTAPP parameters and passed `Python` via the `rpy2` interface to the R functions used for prediction. TAP predictions at a given proteasomal cleavage site are averaged to make a final prediction, but proportional values for the proteasomal fragments at a location can also be taken into account to provide a more representable score. This method is explained later.

### 2.1.3 MHC Affinity and Immunogenicity Prediction

#### MHC Ligand Affinity

Of all the prediction tools available for steps in APP, none exist in greater variety than MHC ligand predictors. Some tools are only available by request or online, while others are available in offline versions. Predictors with only online versions available include SVMHC (Dönnes and Elofsson, 2002), BIMAS (Parker *et al.*, 1994), MAPPP (Hakenberg *et al.*, 2003) and many others.

Table 2.4: TAP ligands are generated from the proteasomal cleavage predictions. The orange coloured residues are cleavage sites (towards the C-terminal end). Only ligands of greater than seven and smaller than sixteen are extracted. The three TAP ligands extracted here are shown.

|                                   |  |
|-----------------------------------|--|
| <b>Initial sequence</b>           | ANNGEDATAGLTHMMIWHSNLPRFKLMV   |
| <b>Proteasomal Cleavage Sites</b> | ANNGEDATAGLTHMMIWHSNLPRFKLMV   |
| <b>TAP Ligands</b>                | <pre> -----ATAGLTHMMIWHSNL----- -----GLTHMMIWHSNL----- -----THMMIWHSNL----- </pre> |

Offline methods include NetMHC (Buus *et al.*, 2003, Nielsen *et al.*, 2004a, Lundegaard *et al.*, 2008) and matrix based methods such as MHCSMM (Tenzer *et al.*, 2005). NetMHC has been tested on numerous occasions as one of the, if not the best, MHC ligand predictors (Trost *et al.*, 2007, Lin *et al.*, 2008). The offline version is a Python script that is designed to run from the command line, however a few modifications allowed the author to directly include it as a module of Fortuna. The can take a list of peptides of lengths 8-11 as input. The word lists are generated in the same way as demonstrated in Section 2.1.1 on page 26. The output is stored in Python dictionaries as peptide lists with the appropriate IC50 value for the requested HLA allotype.

### MHC Ligand Immunogenicity

To the author's knowledge, there is only one immunogenicity predictor available, namely POPI (Tung and Ho, 2007). The tool can take a peptide of any length as input and predicts immunogenicity as four levels: None, Little, Moderate, High. Each higher level represents one  $\log_{10}$  more spot forming units, represented as 0, 1, 2 and 3 for None, Little, Moderate and High predictions. The offline version of the tool was recreated according to criteria in the article. The method utilizes an SVM for predictions. Using the criteria listed, the input peptide is encoded in Python then passed to the libsvm module for the R statistical program where the actual predictions are made. The author cross-checked the locally produced POPI and the version available online, with the result being that 100% of the input peptides were predicted with the same levels of immunogenicity. The input peptides that are used for MHC prediction are also passed to POPI for immunogenicity predictions.

#### 2.1.4 Combining Pathway Predictions

Combining all the predictions in the APP is by no means novel and various tools exist that do integrate proteasomal cleavage, TAP affinity and MHC affinity. NetCTL integrates NetChop, TAPSMM and NetMHC (Larsen *et al.*, 2005) while MHC-Pathway integrates ProteaSMM,

TAPSMM and MHCSMM. The method here differs both in that immunogenicity prediction is taken into account as well as the way TAP predictions are integrated into the simulated pathway. The assumption of proportionality is key to the method presented here. For instance, if two TAP fragments of near equal affinity exist, the one with the higher amount will be more readily transported across the ER. This assumption is also held for MHC affinity where the availability of TAP ligands will determine how many MHC ligands will be available to bind to MHC. Finally, how many potential epitopes will be displayed on the cell surface for interaction with appropriate TCRs.

### Combining TAP and Proteasomal Cleavage Predictions

As stated before, the ability of an MHC ligand precursor to be transported by TAP into the ER is dependent on the TAP fragments that contain said MHC ligand. In turn, these fragments are dependent on the proteasome to be formed. Since proteasomal cleavage is a quantitative process and fragment formation depends on the cleavage promiscuity of the flanking regions, it stands to reason that both the TAP affinity and level of these fragments play a role in determining the rate of transport by TAP into the ER. The other methods take the average  $\log_{IC50}$  of all the TAP fragments of a specific range of lengths that occur upstream from a specific cleavage site. Here, the relative amount of the fragments and their TAP affinity are summed to obtain a single, weighted  $\log_{IC50}$  score. The theory comes from the formula for the dissociation constant as demonstrated in Equation 2.4.

$$K_{D_L} = \frac{[T][L]}{[TL]} \quad (2.4)$$

Where  $K_{D_L}$  is the dissociation constant for the ligand, L, [L] and [T] is the concentrations of the TAP molecule and free ligand respectively whereas [TL] is the concentration of the TAP-ligand complex. The IC50 value is a good approximation of the  $K_{D_L}$  value and by rearranging the formula we can see that the level of [TL] is influenced by [L] and  $[K_{D_L}]$ , assuming the concentration (i.e. availability) of TAP molecules remain the same.

$$[TL_{Relative}] = \frac{[T][L]}{IC50} \quad (2.5)$$

The terms are the same as described for Equation 2.4, with IC50 now substituting  $K_{D_L}$ . The assumption is that for all ligands the same amount of TAP molecule is available for binding. This is not true on a very technical level, as the first ligands to bind to the TAP molecule will obviously decrease its availability. The author assumes, however, that transport of bound ligands are rapid enough for this factor to be trivial.

The pseudo concentration of the ligand,  $L$ , can be predicted as the fraction produced by

Equation 2.2 on page 25. For example, if two solutions contained the same amount of membrane bound TAP molecule (2500 nM) and identical ligands with an IC<sub>50</sub> value of 300nM, but at different relative concentrations (0.6 and 0.1 respectively) were added, their [TL] values would be:

$$[\text{TL}_{0.6}] = \frac{[2500\text{nM}][0.6]}{300\text{nM}} = 5.00$$

$$[\text{TL}_{0.1}] = \frac{[2500\text{nM}][0.1]}{300\text{nM}} = 0.83$$

Essentially, six times the amount of ligand was bound in solution one than was in solution two. Note that the  $K_D$  term was supplanted by the IC<sub>50</sub> value. The goal here is to obtain an IC<sub>50</sub> value for the total amount of fragments associated with a particular epitope. Since [T] remains constant, [L] can be used to directly change the value of the IC<sub>50</sub> value, thus making the value of [TL] inversely correlated to the value of IC<sub>50</sub>. Taking the same value for the solutions used earlier with the adjusted IC<sub>50</sub> values, the same answers are obtained.

$$[\text{TL}_{0.6}] = \frac{[2500\text{nM}]}{300\text{nM} \times \frac{1}{0.6}} = \frac{2500\text{nM}}{500\text{nM}} = 5.00$$

$$[\text{TL}_{0.1}] = \frac{[2500\text{nM}]}{300\text{nM} \times \frac{1}{0.1}} = \frac{2500\text{nM}}{3000\text{nM}} = 0.83$$

To explain the rationale of other researchers behind using averaged  $\log_{\text{IC}_{50}}$  values, we again turn to Equation 2.5. To get the average concentration of all the TAP ligands in a solution is a simple matter of averaging the concentrations. Since Equation 2.5 predicts relative concentrations of [TL] and we know that IC<sub>50</sub> is inversely correlated to [TL], the reciprocal individual IC<sub>50</sub> values for the TAP ligands can be averaged.

$$[\text{TL}]_{\text{Avg}} = \frac{1}{\text{IC}_{50_{\text{Avg}}}} = \frac{\sum_{i=1}^n \frac{1}{\text{IC}_{50_i}}}{n} \quad (2.6)$$

$$[\text{TL}]_{\text{Avg}_w} = \frac{1}{\text{IC}_{50_{\text{Avg}_w}}} = \sum_{i=1}^n \frac{[\text{L}_i]}{\text{IC}_{50_i}} \quad (2.7)$$

Where  $\text{TL}_{\text{Avg}}$  is the average relative concentration of the TAP-Ligand complex,  $\text{IC}_{50_{\text{Avg}}}$  the average IC<sub>50</sub> value and  $n$  the amount of fragments for which the average is calculated. The difference between Equations 2.6 and 2.7 is that in the latter, fragment concentrations are also taken into account and is the method used here and as a result, the average is implied by summation. Of course, if the sum of the  $[\text{L}_i]$  values do not equal one because the C-terminal end from which the TAP ligands originate have a proteasomal score of less than 1.0, the score produced

Table 2.5: Relationship between Scores and Amount

| Score                | Relationship to Amount          |
|----------------------|---------------------------------|
| Proteasomal Cleavage | Directly proportional to amount |
| TAP IC50             | Inversely correlated to amount  |
| MHC IC50             | Inversely correlated to amount  |

by Equation 2.7 will in the majority of cases be less than that of Equation 2.6. Therefore, to negate this ‘total proteasomal’ effect, the value can be divided by the sum of the  $[L_i]$  values to give the IC50 with respect to the relative frequencies of the TAP ligands among themselves. To illustrate all these points, the values of  $\frac{1}{IC50_{Avg}}$  will be compared for three ligands with a IC50 values of 500nM, 300nM, 700nM and proteasomal probabilities of 0.2, 0.4, 0.1 respectively.

$$[TL]_{Avg} = \frac{1}{IC50_{Avg}} = \frac{\sum_{i=1}^n \frac{1}{IC50_i}}{n} = \frac{\frac{1}{500} + \frac{1}{300} + \frac{1}{700}}{3} = 2.225 \times 10^3$$

$$IC50 = \frac{1}{2.2254 \times 10^3} = \mathbf{444nM}$$

$$[TL]_{Avg_{Pw}} = \frac{1}{IC50_{Avg_{Pw}}} = \sum_{i=1}^n \frac{[L_i]}{IC50_i} = \frac{0.2}{500} + \frac{0.4}{300} + \frac{0.1}{700} = 1.876 \times 10^3$$

$$IC50_{Pw} = \frac{1}{1.8762 \times 10^3} = \mathbf{533nM}$$

$$IC50_w = IC50_{Pw} * \sum_{i=1}^n [L_i] = 533 \times 0.7 = \mathbf{373nM}$$

Where IC50 is the IC50 value by directly averaging the reciprocal IC50 results,  $IC50_{Pw}$  is the IC50 value of taking proteasomal cleavage in its entirety into account and  $IC50_w$  is the IC50 value when taking just the relative frequencies of the fragments into account. The above example shows just how much the relative frequencies of the TAP ligands can have an effect on the final IC50 value.

### Combining All Predictions

The final prediction before inclusion of immunogenicity, is to predict the level of a potential MHC ligands relative to other MHC ligands. All the relationships between score and amount are known and again summarised in Table 2.5.

Each step is dependent on the output from the previous step and as such, each preceding score can be treated as a coefficient. The proteasomal scores are used directly and IC50 values for TAP and MHC predictions are in their reciprocal form. The final is expressed as a  $\log_{10}$

value of the product of all the scores. The log values of reciprocal forms of IC50 values may cause the final score to be negative so the score is also adjusted by subtracting the threshold values set for each step. For example, the  $\log_{10}$  value of  $\frac{1}{300}$  is  $-2.48$ ; by subtracting a threshold value of  $\log_{10}\frac{1}{500}$ , the value is now 3.74. In this case, it is clearer which ligands are likely to bind to MHC, by taking every MHC ligand to have a score of greater than or equal to zero. A potential ligand that matches the threshold at each step has a score of zero. Instead of multiplying the scores and then calculating the  $\log_{10}$  value of it, the individual  $\log_{10}$  scores for each step are summed to allow flexibility in terms of choosing which steps to include in calculation of the final score.

$$S_l = (\log_{10}P_l + \log_{10}\frac{1}{TAP_l} + \log_{10}\frac{1}{MHC_l}) - b \quad (2.8)$$

Where  $S_l$  is the final score for the MHC ligand in question,  $P_l$  is the proteasomal score at the cleavage point,  $TAP_l$  is the adjusted IC50 value for all the TAP ligands, but only taking relative frequencies of the TAP ligands into account,  $MHC_l$  the IC50 value for the MHC ligand and  $b$  is the  $\log_{10}$  value of product of all the thresholds for proteasomal, TAP and MHC prediction;  $b$  is calculated by Equation 2.9.

$$b = \log_{10}(P_{\text{threshold}} \times \frac{1}{TAP_{\text{threshold}}} \times \frac{1}{MHC_{\text{threshold}}}) \quad (2.9)$$

Where  $P_{\text{threshold}}$  is the proteasomal threshold score measured in fractional amount, i.e. the actual threshold value divided by the maximum prediction value,  $TAP_{\text{threshold}}$  and  $MHC_{\text{threshold}}$  are the IC50 values of TAP and MHC prediction respectively. The inclusion of any of the scores is optional, but the thresholds used in Equation 2.9 are dependent on the scores utilised in Equation 2.8. If at any point one of the scores has a value of less than or equal to zero before log-transformed, the score is assumed to be infinitely negative. The prediction value of immunogenicity of the MHC ligand can be added to the score directly, since it is already expressed as a  $\log_{10}$  value.

## 2.2 Discovering Potential Self-Epitopes via BLAST

To discover potential self epitopes contained in the input sequences, BLAST is utilised (Altschul *et al.*, 1990). A redundant list of 8-11 mer peptides are built from the input sequences and scanned against a RefSeq database of protein sequences (Pruitt *et al.*, 2005, Pruitt *et al.*, 2007, Pruitt *et al.*, 2009). The rationale behind using RefSeq sequences is that it is a database of curated sequences which reduces redundancy dramatically. Further filtering is performed by excluding records that have the text HYPOTHETICAL or PUTATIVE in the title. To filter out sequences, a list of GI's

are built that meet the exclusion criteria, i.e. they are of Human origin and do not contain **HYPOTHETICAL** or **PUTATIVE** in the title. All the records that do not match the built GI list are extracted and a new reduced version of the database is built. The reduced size of the database also BLAST searches by reducing the number of records to scan. The offline BLAST toolkit, BLAST+, is used to perform all the operations including query searches (Camacho *et al.*, 2009).

### Setting BLAST+ Parameters

The default parameters for BLAST are not optimised for the search of small sequences and adjustments need to be made to them. The recommendation by the BLAST guide to search for small sequences is to:

1. Use the *PAM30* matrix
2. Adjust the word size to 2
3. Use a higher E-value

The scoring matrix is used to measure how well the query sequence matches a testing sequence. The PAM30 matrix is recommended, however the BLOSUM35 matrix is used, because it has already been used in the literature for measuring cross-reactivity and would be ideal for identifying potential self-epitopes (Frankild *et al.*, 2008). The smaller word size will include more sections of a sequence into the search query. One of the blast outputs is an E-value that is, briefly, a measurement of how likely it is for the input query sequence to be the same as the target sequence. The E-value is usually set low to exclude any 'random matches' to a search query. Here, because the search queries are so small, the E-value tends to be naturally higher and the threshold needs to be increased to accommodate potential hits. It is set to 200.

### Measurement of Cross-Reactivity

The method by Frankild *et al.* (2008) is used to measure cross-reactivity between a predicted epitope and potential self-epitope. The method measures the difference between two peptides by utilising the BLOSUM35 matrix. The scores for each residue comparison are tallied and summed. To ensure that the score always falls in the range  $[0, 1]$ , the maximum and minimum comparison scores for each peptide are calculated and the score range scaled. The final comparison score is defined in Equation 2.10.

$$S_{P_1P_2} = \frac{\sum_{i=1}^n B(P_1i, P_2i) - \min(S_{P_1})}{\max(S_{P_1}) - \min(S_{P_1})} \quad (2.10)$$

Where  $S_{P_1P_2}$  is the comparison score of peptide  $P_1$  vs peptide  $P_2$ ,  $P_ni$  the residue of each peptide at position  $i$ ,  $B$  the BLOSUM35 matrix,  $\max(S_{P_1})$  the maximum comparison score, essen-

tially  $P_1$  vs  $P_1$  and  $\min(S_{P_1})$  the minimum comparison score for  $P_1$  calculated by the sum of the minimum value for each residue at all the positions. Even though a BLOSUM matrix is symmetrical, it should be noted that  $S_{P_1P_2} \neq S_{P_2P_1}$  specifically because the values of  $\min(S_{P_1})$  and  $\min(S_{P_2})$  are not the same. Frankild also postulated the idea that weight assignment to each position might augment cross-reactivity measurement. The weights from their research can be included in Equation 2.11.

$$S_{w_{P_1P_2}} = \frac{\sum_{i=1}^n B(P_1i, P_2i)w_i - \min(S_{w_{P_1}})}{\max(S_{w_{P_1}}) - \min(S_{w_{P_1}})} \quad (2.11)$$

Where  $Sw$  is the weighted cross-reactivity score,  $w$  is the weight vector and  $w_i$  the weight assigned at position  $w_i$ . The vector is nine units long and is expanded and contracted for longer and shorter peptide lengths by an averaging procedure.

### Relating Epitopes and BLAST Hits

One peptide word relates to all its BLAST hits, if any. However, since epitopes are not static and can occur with various substitutions, the reverse is also applied, meaning that a BLAST hit also relates to all the variants of an epitope for which it was a hit. The total score for any BLAST hit can be either the average cross-reactivity score to all the potential epitope hits and/or the product of the average score to the sum of the frequencies of all the epitopes across all the input sequences. This augmentation can possibly reveal how important it is for the pathogen in question to retain this potential self-epitope.

## 2.3 Analysis and Visualisation

In this section, it will be shown how the prediction methods of the previous section can be implemented in analysis and visualisation procedures. First, rudimentary visualisation of predicted MHC affinity of epitopes and variants across the input sequence(s). Next, it will be shown how prediction data from multiple sequences can be used to perform frequency and entropy analysis on the data, showing the occurrence of potential epitopes at positions as well as the frequency of the variants that occur there. Following this, is the cluster analysis procedures. Here the development of a method that incorporates both the Frankild score and pathway prediction results to determine distances between two epitopes. Extending from this, is the development of a clustering procedure based on the calculated distances. Finally, it will be shown how the clustering and comparison of clusters can be visualised.



### 2.3.1 MHC Treemap and Density Plots

Selective pressures applied on the sequence of a potential epitope has an inherent effect on its affinity value. Escape mutations include those that will render the epitope unable to bind sufficiently to MHC to elicit a response. Given that multiple variants of a sequence is given, the total variation of MHC affinity can be calculated and visualised through density plots. Furthermore, the frequency of epitopes per HLA allotype is also an important factor in determining for which allotypes immunologically sensitive. Various methods exist to visualise relative frequencies, e.g. bar graphs, pie charts etc. Another effective method is a Treemap. A treemap is a rectangular plot that is made up of smaller rectangles whose area illustrates the input values used for its construction. In this specific case, the frequency of input sequences that contain epitopes of a given length for a given HLA allotype. The regions for each allotype are assigned a unique hue of colour. Furthermore, the colours are assigned a lightness based on the log value of the average of the log IC<sub>50</sub> values for all the potential binders. The procedure to construct the treemap can be summarised as:

1. Tally the epitopes for the queried HLA allotypes and epitope lengths
2. Divide the plot into areas that reflect the total frequency of MHC ligands for an HLA allotype
3. Further sub-divide the areas into areas reflecting the frequency of an MHC ligand at a given position for a particular HLA-allotype

To demonstrate, the potential MHC ligands of the HIV protein, p17, were predicted for HLA allotypes B3501 and A0201 with the ligand lengths restricted to 9 and 10 residues. The treemap is shown in Figure 2.4 on the next page. Specific information is also displayed in two tables. The first table, 2.6 on the following page, contains information on the frequency and average IC<sub>50</sub> values for MHC ligands occurring at specific positions. The second Table, 2.7 on page 41, contains a summary of the total information of MHC ligands of the queried lengths across all sequences for each allotype. The intention is not to give a full detailed analysis of MHC ligands, merely a summarized version to determine which positions could be key in epitope analysis. Detailed analysis of MHC ligands in terms of variability in conjunction with frequency is described in Section 2.3.1 on page 42. Please note that the discrepancies between Tables 2.6 and 2.7 are due to omission of some of the values that existed in the original table from which Table 2.7 was constructed.

Density plots are created from the MHC affinity predictions. Density plots visually represent the frequency of sequences containing an MHC ligand at a specific position at and above a range of log<sub>IC<sub>50</sub></sub> values. These plots are generated for the peptides at each position for which at

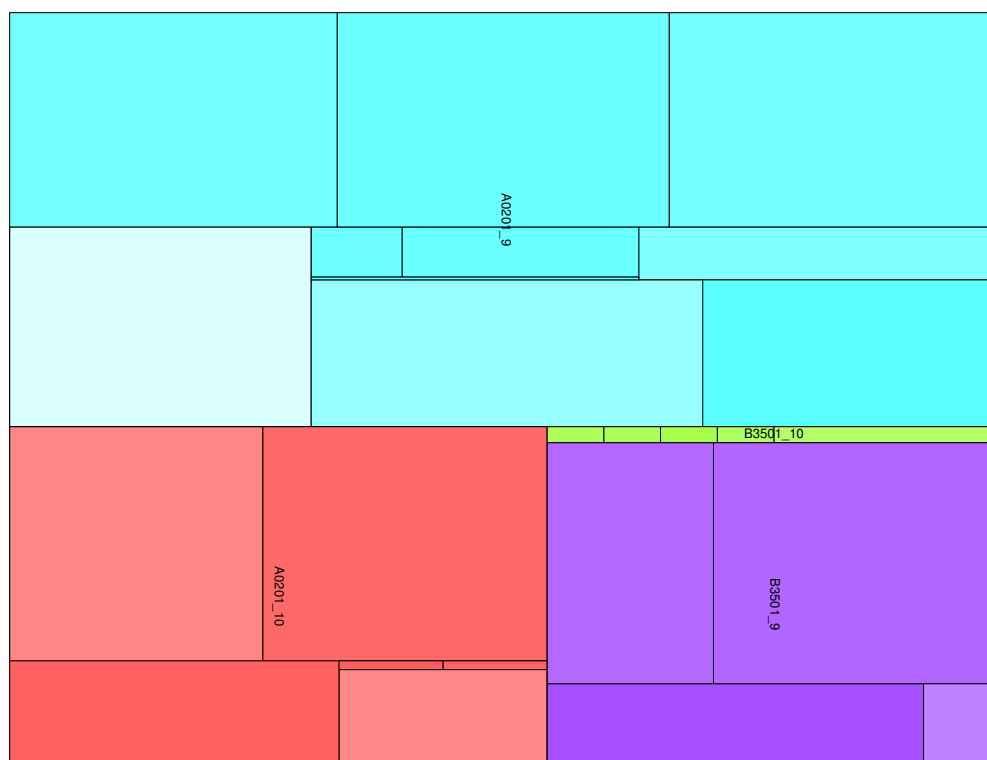


Figure 2.4: Example of an MHC Treemap. The graph is divided into sectors. The hue represents an allotype. The area of the sector corresponds to the total MHC ligands of that particular allotype. The sector is further sub-divided into smaller areas representing the allotype frequency at a specific position. Each smaller block represents the frequency of sequences that contained predicted MHC ligands at a single position. The lightness of the hue represents the reciprocal average IC<sub>50</sub> value.

least one MHC ligand was predicted for a given allotype. Again, sequences of the HIV protein, p17, is used as an example and displayed in Figure 2.5 on the next page.

Table 2.6: Epitope Counts per Allotype at position in Sequences.

| Allotype | Length | Position | Avg MHC | Count | Fraction |
|----------|--------|----------|---------|-------|----------|
| A0201    | 10     | 20       | 199     | 73    | 0.94     |
| A0201    | 10     | 37       | 82      | 65    | 0.83     |
| A0201    | 10     | 91       | 233     | 1     | 0.01     |
| A0201    | 10     | 92       | 270     | 1     | 0.01     |
| A0201    | 9      | 34       | 171     | 5     | 0.06     |
| A0201    | 9      | 36       | 6       | 66    | 0.85     |
| B3501    | 10     | 60       | 245     | 1     | 0.01     |
| B3501    | 10     | 67       | 450     | 1     | 0.01     |
| B3501    | 9      | 60       | 92      | 7     | 0.09     |
| B3501    | 9      | 85       | 206     | 44    | 0.56     |

Table 2.7: Epitope Counts per Allotype at position in Sequences.

| Allotype | Length | Count | Avg Count per Position |
|----------|--------|-------|------------------------|
| A0201    | 10     | 200   | 33                     |
| A0201    | 9      | 450   | 45                     |
| B3501    | 10     | 8     | 1                      |
| B3501    | 9      | 161   | 40                     |

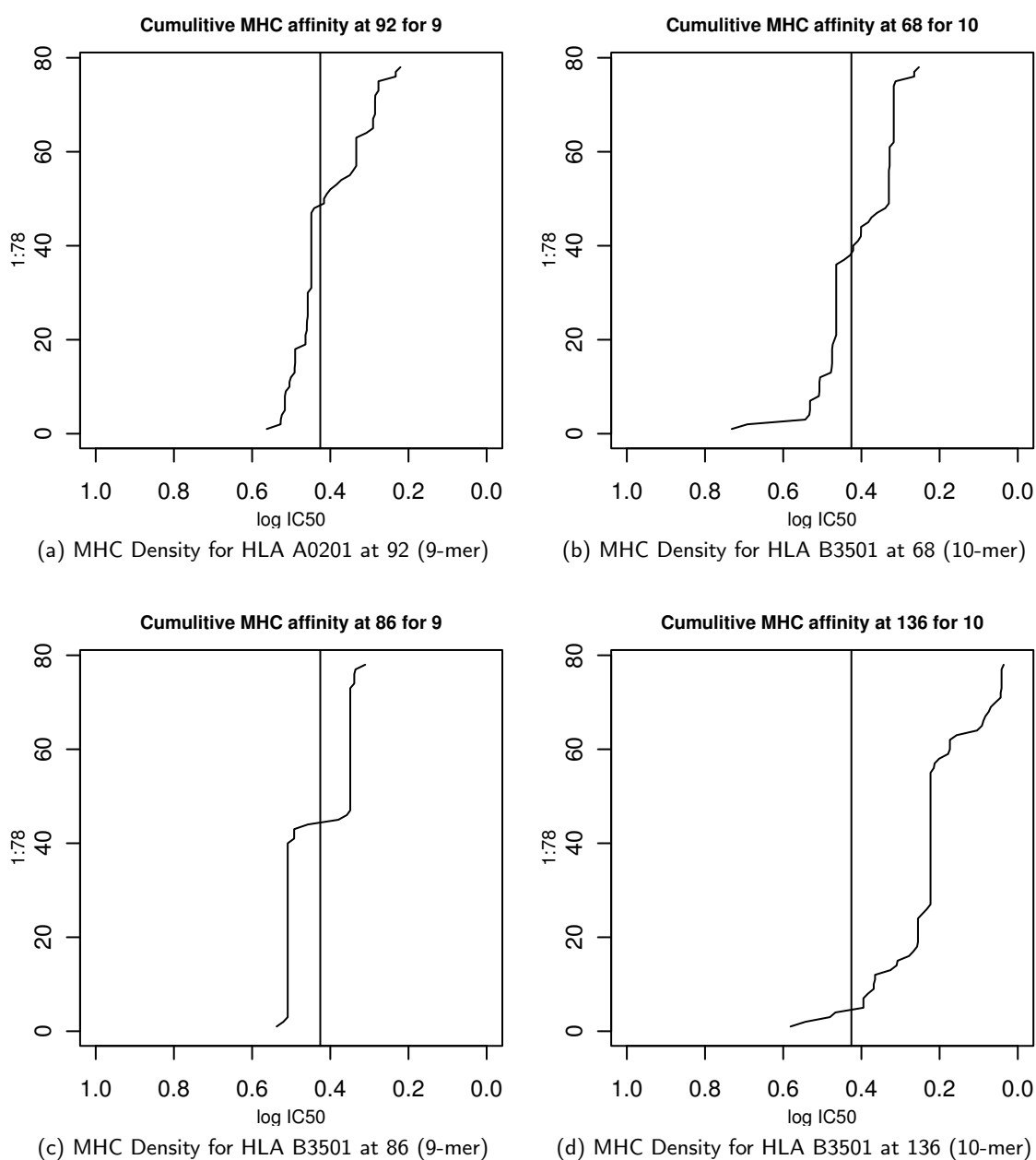


Figure 2.5: MHC Density Plots for HLA Allotypes A0201 at positions 92 and 68 and B3501 at 86 and 136. The y-axis is the cumulative frequencies of sequences containing MHC ligands with at most the  $1 - \log_{50000} IC_{50}$  values represented by the x-axis. The vertical line is the preset threshold that constitutes MHC binding, in this case represented by  $1 - \log_{50000} 500$ .

### 2.3.2 Entropy and Frequency Analysis

Difference between sequences of the same protein can account for immunological escape. Factors like MHC affinity and immunogenicity can directly be influenced if a mutation exists within the site where an MHC ligand exists. Mutations that affect MHC affinity negatively may negate the ability of a ligand to bind to MHC and thus never be under the scrutiny of the immune system. On the other hand, there can be mutations that negatively affect the interaction between MHC-ligand and the TCR. Thus, sequence variability measured against epitope frequency at a particular range of positions and the sequence variability of predicted MHC ligands can shed light into immunological escape mechanisms.

#### Entropy as a Measurement of Sequence Variability

The Shannon Entropy score is a measurement of variability of a sequence of characters (Shannon, 1948). It utilizes the frequency and diversity of characters in a word to assess the variability. This is best explained by a direct example. The words **GOOGOL** and **SPLASH** have different entropies since **GOOGOL** has more frequently occurring letters than **SPLASH**. Calculation of their respective Shannon entropies is accomplished through Equation 2.12.

$$H(X)_b = - \sum_{i=1}^n p(x_i) \log_b p(x_i) \quad (2.12)$$

Where  $H(X)$  is the entropy of a particular word,  $n$  the length of the word,  $x_i$  a particular character in the word,  $p(x_i)$  the frequency of the character and  $b$  the log base used. Applying the formula to the words **GOOGOL** and **SPLASH** using a log base of 2, the following values are obtained:

$$\begin{aligned} H(\text{GOOGOL}) &= -[p(G)\log_2(G) + p(O)\log_2(O) + p(L)\log_2(L) + p(E)\log_2(E)] \\ &= -[(0.33)(\log_2 0.33) + (0.5)(\log_2 0.5) + (0.17)(\log_2 0.17) + (0.17)(\log_2 0.17)] \\ &= 1.90 \end{aligned}$$

$$\begin{aligned} H(\text{SPLASH}) &= -[p(S)\log_2(S) + p(P)\log_2(P) + p(L)\log_2(L) + p(A)\log_2(A) + p(H)\log_2(H)] \\ &= -[(0.33)(\log_2 0.33) + (0.17)(\log_2 0.17) + (0.17)(\log_2 0.17) + \\ &\quad (0.17)(\log_2 0.17) + (0.17)(\log_2 0.17)] \\ &= 2.27 \end{aligned}$$

It is shown that SPLASH having more variable characters, and as a consequence lower average frequencies, does indeed have a higher  $H(X)$  value of 2.27 as opposed to GOOGOL with a value of 1.90. Since amino acids can be represented by single characters, Shannon entropy can also be applied to measure variability. The terms of the Equation are redefined as:

- $H(X)$  – The entropy of all the amino acids occurring across all sequences at aligned position  $X$ . The score is rescaled so that the maximum possible entropy is 1.0 and the minimum entropy 0.0
- $x$  – A vector containing the frequency of occurring amino acids
- $p(x_i)$  – The frequency of the amino acid at the vector position  $i$

Table 2.8 provides an example of entropy measurements across the positions of the sequence KQIMKQLQP and variants of it. The entropy is 0.00 for most positions, except positions 4, 5 and 7 which have values of 0.66, 0.66 and 0.31. The lower row of numbers is the windowed average version of entropy. These values are measured by obtaining the average of the entropy values at, and 8 positions upstream from it. If the N-terminal end of the sequence is less than 9 amino acids from the measured entropy value, the window size is shortened according to how far the N-terminal is. The averaging procedure is done give a single score across the entire range of positions that contain an MHC ligand. Of course, for differently sized MHC ligands, a different window size is used.

Table 2.8: Example of Shannon entropy calculation for five sequences that are nine amino acids long. The higher variable regions have higher entropy. The entropy score is range scaled for the amount of sequences. The smoothed score is calculated by window averaging.

| Position         | 1    | 2    | 3    | 4    | 5    | 6    | 7    | 8    | 9    |
|------------------|------|------|------|------|------|------|------|------|------|
| Seq 1            | K    | Q    | I    | M    | K    | Q    | L    | Q    | P    |
| Seq 2            | K    | Q    | I    | M    | K    | Q    | L    | Q    | P    |
| Seq 3            | K    | Q    | I    | I    | Q    | Q    | L    | Q    | P    |
| Seq 4            | K    | Q    | I    | I    | Q    | Q    | L    | Q    | P    |
| Seq 5            | K    | Q    | I    | L    | A    | Q    | I    | Q    | P    |
| Entropy          | 0.00 | 0.00 | 0.00 | 0.66 | 0.66 | 0.00 | 0.31 | 0.00 | 0.00 |
| Entropy Smoothed | 0.00 | 0.00 | 0.00 | 0.16 | 0.26 | 0.22 | 0.23 | 0.20 | 0.18 |

### MHC Ligand Frequency Calculation

To determine the frequency of MHC ligands at specific positions, the sequences are first screened for MHC ligands of a set of lengths and associated with a set of HLA allotypes and the frequencies of sequences containing MHC ligands at a certain position. This is best illustrated by an example

Table 2.9: MHC Ligand Frequency Calculation.

|           | Sequence                                     |
|-----------|--|
| Seq 1     | RLRPGGKKT <u>YMLKHLV</u> WASRELERFALNPGLLETA |
| Seq 2     | RLRPGGKKT <u>YMLKHLV</u> WRSRELERFALNPGLLETA |
| Seq 3     | RLRPGGKKT <u>YMLKHLV</u> WRSRELERFALNPGLLETA |
| Frequency | 0000000334444444110000000222222220           |

shown in Table 2.9. Subsections of three sequences of HIV protein p17 were analyzed for MHC ligands. MHC ligands are shaded to either **X** or **X**. The underlined amino acid indicates the C-terminal of the MHC ligand. To deal with overlapping sequences, the frequencies are determined across the entire epitope length. In Seq 1 there is a region of overlap between two MHC ligands, indicated by the **X** colour. Reading the frequencies, we can see that the ligand YMLKHLVWA is also taken into account when calculating the frequencies. Looking at the ligand ALNPGLLET, we can see that there are no overlapping ligands, so only those ligands are taken into account for frequency calculation. In this specific example, only MHC ligands of length 9 and associated with HLA allotype A\*0201 are shown. In practice, all the MHC ligands for all the queried HLA allotypes will be tallied together.

### Correlating Entropy and Frequency

The entropy score increases as the variability of amino acids increase. This, in conjunction with frequency of sequences containing an MHC ligand at a specific or range of positions can be used to predict whether the mutations within the MHC ligand regions have a significant effect on MHC binding by determining correlation between frequency and entropy as demonstrated in a previous study (Yusim *et al.*, 2002). The procedure used by Yusim is useful for determining total effect of mutations on *all* regions containing MHC ligands. However, since MHC affinity alone cannot account for immunogenicity of a ligand, the procedure is modified here by estimating entropy/frequency correlations at localised positions in a sequence. For both entropy and frequency, window smoothed versions of the scores are used and correlation is measured by Spearman's method. The rationale behind smoothing the scores is to determine correlation based on change in the two parameters, i.e. does the change in entropy have an effect on frequency of MHC ligands. Spearman's rank correlation is used in case the data is not normally distributed, which would be a prerequisite of Pearson's correlation.

To calculate which local areas to test, the positions of consecutive high and low peaks of the frequencies are determined. This is accomplished by using a smoothing function in the R language. A minimum length of typically 15 is imposed on correlation testing. An example of the smoothing procedure and how the minimum length is imposed is shown in Figure 2.6 on the following page.

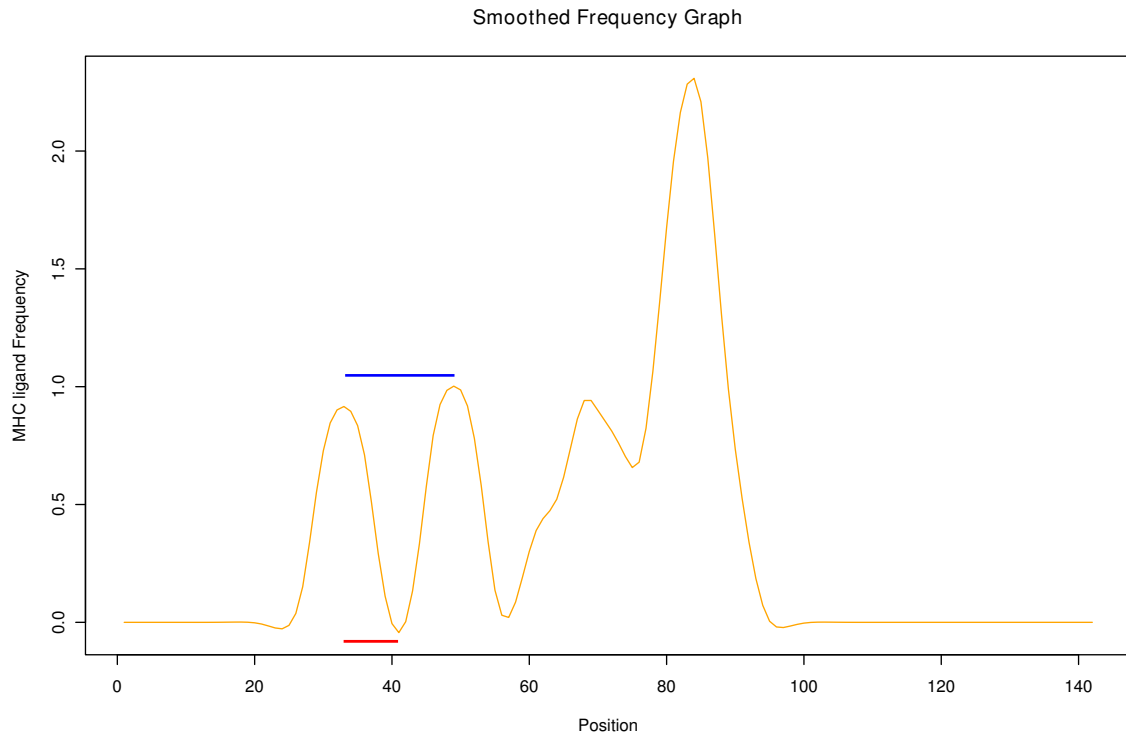


Figure 2.6: Example of smoothing the MHC ligand frequencies. The orange line is the smoothed version of the frequencies. The peaks of the graph are determined numerically. Since limitations is placed on the minimum length of a region to be tested for frequency and entropy correlation, the length between two peaks may be too small, as indicated by the red line. The solution is to extend the testing region to the next peak as indicated by the blue line.

Finally the correlations are plotted together with frequency and entropy. Both positive and negative correlations are included, though negative correlations are more useful in making conclusions about the entropy/frequency relationship. An example of entropy vs frequency output is given in Figure 2.7.

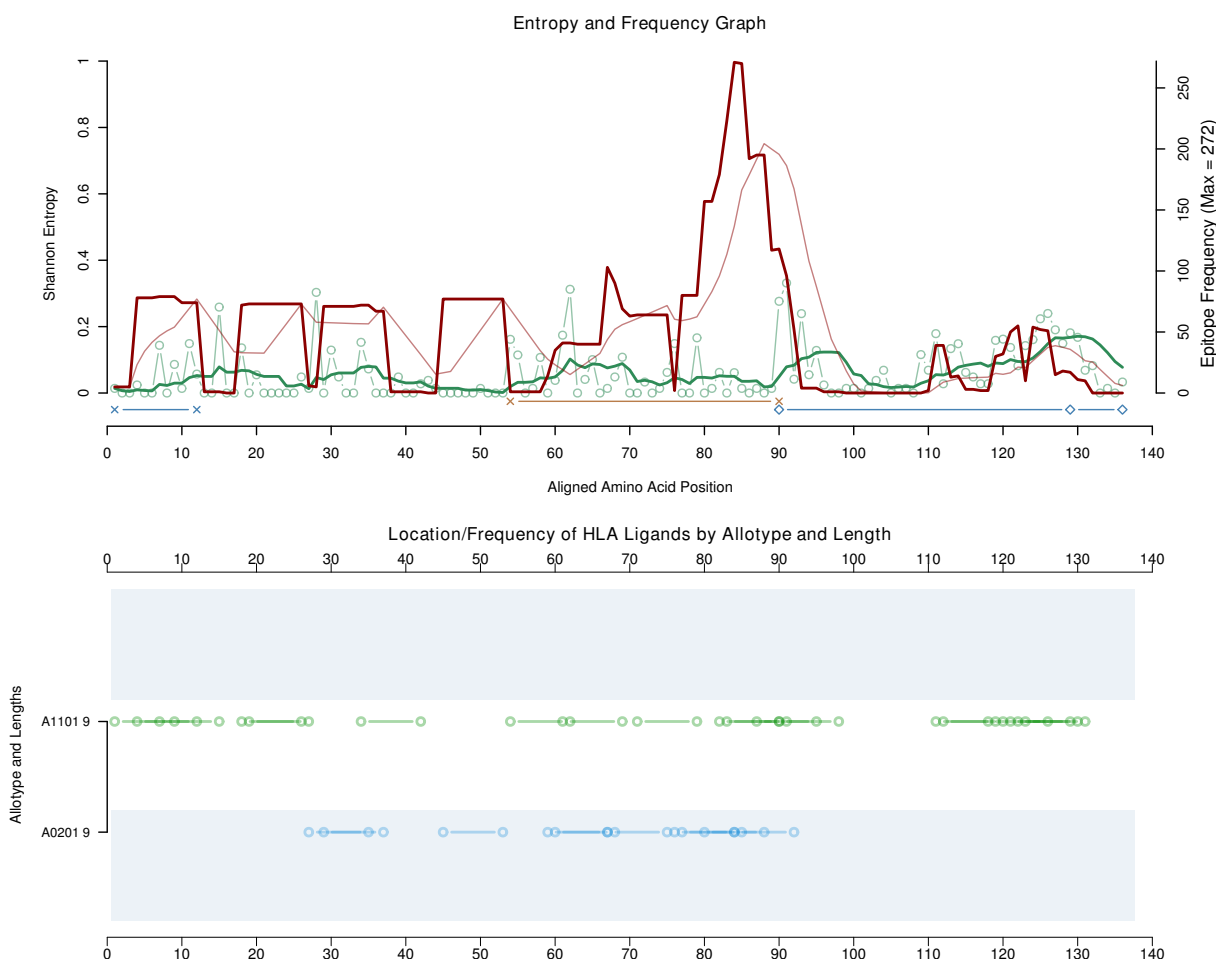


Figure 2.7: The Figure shows actual output from entropy vs frequency analysis of HIV p17. The keys of the graphs have been excluded. The stepping red line is the actual measured, overlapping frequency of MHC ligands at their respective positions. The smoother red line is the window averaged frequency. The thin, stepping green lines indicate entropy while the thicker, smoother green lines represent the window averaged entropy. Below the graph are lines indicating significant local correlations, orange and blue representing negative and positive correlations respectively. The tick marks at the edges of the lines indicate the boundaries of local correlation. The bottom graph is a map of MHC ligand occurrence for the different allotypes queried. The frequency of a ligand is determined by how opaque it is. In both the bottom and top graph, the x-axis represent aligned position of the sequences.

### SeqLogos for MHC ligands

Since immunogenicity isn't granted to all MHC ligands, it is important to observe mutations that do occur in predicted ligands that may have an effect on immunogenicity. As a visual abstraction, the local version of the SeqLogo package is employed (Crooks *et al.*, 2004) to survey predicted MHC ligands for mutations. An example is given in Figure 2.8 on the next page. The SeqLogo output can be considered of a two-dimensional nature. In addition to providing encountered amino acids at specific positions within the region of the protein sequence and their relative frequency, the total height of the combined amino acids also reveal how conservative a position is within a sequence.

To fully illustrate MHC ligand variants, the results are displayed in a table with HLA allotype, position within the sequence and the frequency of sequences containing an MHC ligand is also



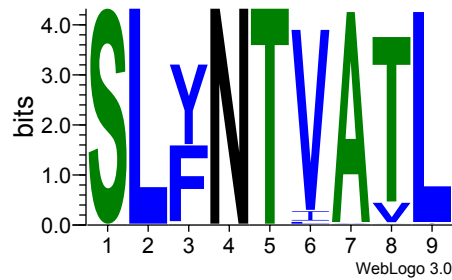


Figure 2.8: The well known SLYNTVATL MHC ligand that occurs on aligned position 85 of the HIV protein p17 has many variants associated with it. The height of the letters representing the amino acids reveal the log version of their frequency within the tested sequences. A notable mutation is seen in position 3, with the substitution of tyrosine for phenylalanine.

| Position | Count | Seqs  | SeqLogo |
|----------|-------|---|---------|
| A0201    |       |   |         |
| 84       | 127   | Hide<br>SLFNTVATL (55)<br>..Y..... (52)<br>.....V. (4)<br>..Y..I..V. (3)<br>..Y...V. (3)<br>..Y..L..V. (2)<br>.....I..V. (1)<br>..Y..I... (1)<br>.....I... (1)<br>..H..... (1)<br>..Y....A. (1)<br>..S..... (1)<br>..Y..IVV. (1)<br>..Y.....I (1) |         |

Figure 2.9: MHC ligand variability example

included. By including frequency, it can also be deduced whether immuno-evasive mutations are more partial to eliminating MHC binding potential or MHC-ligand-TCR interaction. A table example of SLYNTVATL and its variants is given in Figure 2.9. Only a single entry is given, but it can be seen that the table is divided into five parts. Firstly, HLA allotypes are grouped together, the first column indicates position, the second the frequency, the thirist shows the most frequently occurring MHC ligand, and then substitution variants of it. The positions of the sequence variants that contain identical amino acids are replaced by a period to easily accentuate the positions that *do* contain different amino acids. The last column contains the SeqLogo.

It may be asked why not all the sequences at a given position are included to give a full spectrum of mutations. The reason is simply that mutations that directly effect immunogenicity without great influence of MHC affinity are isolated.

### 2.3.3 Epitope Info

Section 2.1.4 on page 32 describes how prediction scores are combined. The next task is to display these predictions in a sensible manner. The predictions for proteasomal cleavage, TAP

affinity, MHC affinity and immunogenicity are all individually important as well as combinations of the scores. Having said that, displaying all the permutations will not reveal enough relevant information. What complicates matters further is especially proteasomal and TAP predictions when taking into account the protein can be processed by either the constitutive or immunoproteasome. Therefore only certain combination forms of predictions are deemed useful. They are listed in Table 2.10 on the following page.

The scores are computed and displayed in a table. An example of the data displayed is shown in Table 2.11 on page 50. All the scores are threshold-adjusted  $\log_{10}$  values of the original score. The rationale behind this is so reveal which scores make up larger combination scores like those that take all predictions into account (MI.PTc and MI.PTi in the table). The threshold-adjustment is essentially subtracting the cutoff value for a prediction from the predicted score. The resultant score that doesn't make the cut-off will be less than zero. For clarification reasons, and to not penalize combination scores too much by one prediction, values below the threshold are adjusted to match the threshold, essentially not contributing to the final score.

The individual epitope scores can prove to be very effective, however it is also useful to see the total score for an individual sequence based on its predicted immunological prowess. A more summarised form of the epitope prediction is utilised, only taking the total prediction results into account, i.e. MIPTc and MIPTi (see Table 2.10 on the following page). The sum of these scores are obtained for all the predicted epitopes in a sequence. To ensure that the score isn't biased towards sheer amount of predicted epitopes available, an average result is also obtained. Finally, the amount of epitopes are counted for the queried HLA allotypes and ligand lengths and combined with the scores are displayed in a table.

#### 2.3.4 Cluster Analysis

Another way in which the prediction results for potential CTL epitopes can be utilised is comparing sequences based on their epitope profiles. Finding similar patterns of CTL epitopes could prove pivotal in, for instance, vaccine design. Phylogenetic methods have been developed to group proteins based on their amino acid sequence. Since mutations within an epitope of a particular HLA or HLA supertype is mostly independent from other epitopes that occur within a sequence, it is prudent to provide a method that computes distances between sequences based on only these predicted epitopes. With appropriate distance measures, hierarchical clustering procedures can be employed to group immunologically similar proteins together as well as revealing immunologically distant groups.

Table 2.10: Pathway Prediction Combination Criteria. Note that 'c' and 'i' are abbreviations for constitutive and immunoproteasome.

| Score   | Abbreviation     | Rationale  |
|---|------------------|--|
| Proteasomal Cleavage  | Pi or Pc         | Reveals the initial amount of the original peptide from which the MHC ligand can be extracted                                    |
| TAP affinity  | Ti or Tc         | Shows the average TAP affinity, which indicates the amount of fragment available for MHC loading                                 |
| MHC affinity  | M                | Indication of MHC-peptide availability on the cell surface   |
| Immunogenicity  | I                | Indication of the immunostimulating potency of the MHC-ligand  |
| Proteasomal Cleavage and TAP affinity                               | PTc or PTi       | Reveals in relative terms how much of the initial peptide will be available for MHC loading                                      |
| Proteasomal Cleavage, TAP affinity and MHC affinity                 | M.PTc or M.PTi   | Reveals in relative terms how much of the peptide will be available on the cell surface  |
| Proteasomal Cleavage, TAP affinity, MHC affinity and Immunogenicity | MI.PTi or MI.PTc | The total stimulatory value of the MHC ligand  |
| MHC affinity and Immunogenicity                                     | MI               | How well the peptide can stimulate the immune response based on innate ability and relative amount available on the cell surface |

### Calculating Sequence Distances

Distances between sequences are calculated by the difference in their predicted epitope repertoires. To accomplish this, the epitopes are mapped to their aligned position within the sequences they are found and epitope sequence comparisons based on the Frankild method are done. An example of immunological comparison between two sequences,  $P_1$  and  $P_2$  based on predicted epitopes for HLA A0201 and HLA B3501 is shown in Table 2.13. One factor that is immediately apparent is the inequality between the scores when comparing  $P_1$  to  $P_2$  and  $P_2$  to  $P_1$ . This is partly because the Frankild score is asymmetrical and partly because of the absence of an A0201 ligand in  $P_2$  that occurs in  $P_1$  at position 54. The absence of a predicted epitope in one sequence equates to a distance of 1 when comparing it with a sequence that does, in fact, contain it. The sum of the scores indicate that  $P_1$  is closer to  $P_2$  than  $P_2$  is to  $P_1$ .

This rather simplistic comparison does not yield unambiguous scores. The immunologically stimulatory effect of the epitopes are not taken into account. For instance, if the predicted epitope ALNPGLETA is not very immunologically active, the exclusion of it from  $P_2$  is not that significant and the distance from  $P_2$  to  $P_1$  should be reduced. The comparison score can be further

Table 2.11: Listed are epitopes by allotype, length and position of the C-terminal amino acid. The abbreviations of some of the headings are explained by their individual characters: **M** is the MHC affinity prediction, **I** is the POPI immunogenicity prediction, **P** is proteasome and **T** is TAP where their subscripts **c** and **i** stand for predictions for immuno- and constitutive proteasome respectively. For example **Ti** means TAP predictions for peptides originating from Immunoproteasome digestion. Combined letters mean combination score of the individual predictions. See text for description of the score values.

| Allotype | Position | Length | Sequence   | Count | Pc   | Tc   | Pi   | Ti   | M   | I | MPTi | MPTc | MIPTi | MIPTc | MI   |
|----------|----------|--------|------------|-------|------|------|------|------|-----|---|------|------|-------|-------|------|
| A0201    | 21       | 10     | KLDTWETIRL | 1     | 0.40 | 0.00 | 0.84 | 0.00 | 17  | 2 | 5.11 | 4.78 | 6.11  | 5.78  | 2.47 |
| B3501    | 41       | 9      | HPVWASREL  | 1     | 0.24 | 0.00 | 0.51 | 0.00 | 17  | 2 | 4.79 | 4.46 | 5.79  | 5.46  | 2.47 |
| A0201    | 88       | 9      | NTVAVLYCV  | 3     | 0.01 | 0.00 | 0.20 | 0.00 | 115 | 3 | 3.57 | 3.24 | 5.57  | 5.24  | 2.64 |
| A0201    | 85       | 9      | SLYNTIATI  | 1     | 0.25 | 0.00 | 0.61 | 0.00 | 148 | 3 | 3.38 | 3.05 | 5.38  | 5.05  | 2.53 |
| B3501    | 101      | 10     | IAVRDTKEAL | 1     | 0.12 | 0.00 | 0.64 | 0.00 | 321 | 3 | 3.32 | 2.99 | 5.32  | 4.99  | 2.19 |
| A0201    | 92       | 9      | TLYCVHAGI  | 1     | 0.02 | 0.00 | 0.21 | 0.00 | 168 | 3 | 3.24 | 2.91 | 5.24  | 4.91  | 2.47 |
| A0201    | 92       | 9      | TLYCVHEGI  | 4     | 0.03 | 0.00 | 0.28 | 0.00 | 188 | 3 | 3.20 | 2.87 | 5.20  | 4.87  | 2.43 |
| A0201    | 21       | 10     | KLDTWEKIRL | 12    | 0.36 | 0.00 | 0.83 | 0.00 | 144 | 2 | 4.19 | 3.86 | 5.19  | 4.86  | 1.54 |
| A0201    | 21       | 10     | KLDKWEGIRL | 1     | 0.52 | 0.00 | 0.92 | 0.00 | 158 | 2 | 4.11 | 3.78 | 5.11  | 4.78  | 1.50 |
| A0201    | 21       | 10     | KLDTWERIKL | 2     | 0.34 | 0.00 | 0.82 | 0.00 | 190 | 2 | 4.02 | 3.69 | 5.02  | 4.69  | 1.42 |
| B3501    | 61       | 9      | TPEGCKQIM  | 2     | 0.33 | 0.00 | 0.58 | 0.00 | 34  | 2 | 3.93 | 3.60 | 4.93  | 4.60  | 2.17 |
| B3501    | 61       | 9      | TPEGCRQIM  | 1     | 0.25 | 0.00 | 0.52 | 0.00 | 39  | 2 | 3.88 | 3.55 | 4.88  | 4.55  | 2.11 |
| A0201    | 85       | 9      | SLFNTVAVL  | 3     | 0.49 | 0.00 | 1.04 | 0.00 | 157 | 3 | 2.82 | 2.49 | 4.82  | 4.49  | 2.50 |
| A0201    | 88       | 9      | NTIATLYCV  | 2     | 0.11 | 0.00 | 0.29 | 0.00 | 80  | 2 | 3.80 | 3.47 | 4.80  | 4.47  | 1.80 |

Table 2.12: Epitope Information for Sequences.

| Seq | Title   | A0201 <sub>9</sub> | A0201 <sub>10</sub> | B3501 <sub>9</sub> | B3501 <sub>10</sub> | Total Epitopes | MIPTi | MIPTc | Avg MIPTi | Avg MIPTc |
|-----|---|--------------------|---------------------|--------------------|---------------------|----------------|-------|-------|-----------|-----------|
| 24  | C ZA 2004 SK143B1                             | 4                  | 1                   | 2                  |                     | 7              | 24.62 | 26.93 | 3.52      | 3.85      |
| 31  | C ZA 2003 SK116B1                             | 6                  | 2                   | 1                  |                     | 9              | 28.15 | 31.12 | 3.13      | 3.46      |
| 49  | C ZA 1999 I051M                               | 5                  | 1                   | 3                  |                     | 9              | 27.44 | 30.42 | 3.05      | 3.38      |
| 76  | C ZA — J129M                                  | 6                  | 3                   | 2                  |                     | 11             | 32.36 | 35.99 | 2.94      | 3.27      |
| 7   | C ZA 2000 I217MB                              | 6                  | 3                   | 1                  |                     | 11             | 31.72 | 35.35 | 2.88      | 3.21      |
| 58  | C ZA 1999 I058M                               | 4                  | 1                   | 3                  |                     | 8              | 22.87 | 25.51 | 2.86      | 3.19      |
| 21  | C ZA 2000 C <sub>z</sub> A <sub>1</sub> 069MB | 5                  | 2                   | 3                  |                     | 11             | 31.40 | 34.70 | 2.85      | 3.15      |
| 48  | C ZA 2000 I201M                               | 9                  | 3                   | 1                  |                     | 13             | 36.73 | 41.02 | 2.83      | 3.16      |
| 26  | C ZA 2003 SK023B2                             | 6                  | 2                   | 3                  |                     | 11             | 31.06 | 34.70 | 2.82      | 3.15      |
| 73  | C ZA 2002 2732M                               | 6                  | 3                   | 2                  |                     | 11             | 30.84 | 34.47 | 2.80      | 3.13      |
| 70  | C ZA 2002 2702M                               | 5                  | 2                   | 3                  |                     | 10             | 27.93 | 31.23 | 2.79      | 3.12      |
| 55  | C ZA 2001 2015M                               | 6                  | 3                   | 3                  |                     | 12             | 33.53 | 37.50 | 2.79      | 3.12      |
| 52  | C ZA 2001 I232M                               | 5                  | 2                   | 3                  |                     | 10             | 27.93 | 31.23 | 2.79      | 3.12      |
| 45  | C ZA 2000 I158M                               | 6                  | 3                   | 3                  |                     | 12             | 33.31 | 37.28 | 2.78      | 3.11      |

Table 2.13: Calculation of immunological distance between the sequences  $P_1$  and  $P_2$  based on some of the calculated epitopes for HLA A0201 and HLA B3501. The score  $1 - S_{E_x E_y}(i)$  is the comparison of the epitopes that exist at position  $i$  in sequences  $x$  and  $y$ . The underlined portions in a sequence indicate the mutational site while '-' indicates that no epitope was predicted for the sequence in question for the HLA allotype and position in question.

| Allotype         | Position | Epitope in $P_1$   | Epitope in $P_2$   | Score ( $1 - S_{E_{P_1} E_{P_2}}(i)$ ) | Score ( $1 - S_{E_1 E_2}(i)$ ) |
|------------------|----------|--------------------|--------------------|--|--------------------------------|
| A0201            | 67       | QIM <u>T</u> QLQPA | QIMA <u>Q</u> LQPA | 0.06                                   | 0.06                           |
| A0201            | 54       | ALNPGLLETA         | -                  | 1.00                                   | 0.00                           |
| B3501            | 61       | T <u>P</u> EGCKQIM | T <u>A</u> EGCKQIM | 0.08                                   | 0.12                           |
| Total Difference |          |                    |                    | 1.14                                   | 0.20                           |

augmented by assigning weights to the epitopes based on pathway prediction scores with Equation 2.13. With the weight augmentation, the sequences are now a lot closer immunologically. MHC ligands for HLA A0201 at positions 54 and 67 are predicted to be non-immunogenic. For both cases, the  $C_w$  score is calculated to 0 and it is only the epitopes for HLA B3501 at position 61 that contributes to the immunological distance between the two sequences.

$$C_{w_{P_1 P_2}} = (1 - S_{E_{P_1} E_{P_2}}) \times S_1(E_{P_1}) \quad (2.13)$$

Where  $C_{w_{P_1 P_2}}$  is the distance *from* epitope  $P_2$  to  $P_1$ ,  $S_{E_{P_1} E_{P_2}}$  is the Frankild score and  $S_1(E_{P_1})$  the predicted pathway score. Note that any combination of predictions could potentially be used to calculate  $S_1(E_{P_1})$ , e.g. the MHC affinity and immunogenicity scores. If the weighted score is less than 0, the score is rounded to 0 meaning that any sequential difference when transitioning from  $P_2$  to  $P_1$  is trivial. Careful attention should be paid not to read  $C_{w_{P_1 P_2}}$  as the distance from  $P_1$  to  $P_2$ . The pathway weighted augmented example of Table 2.13 is show in 2.14.

Table 2.14: The same sequences and epitopes are compared as in Table 2.13, but now augmented with the weights of MHC affinity and immunogenicity.

| Allotype         | Position | Epitope in $P_1$   | Epitope in $P_2$   | Score ( $C_{w_{P_1 P_2}}$ ) | Score ( $C_{w_{P_2 P_1}}$ ) |
|------------------|----------|--------------------|--------------------|-----------------------------|-----------------------------|
| A0201            | 67       | QIM <u>T</u> QLQPA | QIMA <u>Q</u> LQPA | 0.00                        | 0.00                        |
| A0201            | 54       | ALNPGLLETA         | -                  | 0.00                        | 0.00                        |
| B3501            | 61       | T <u>P</u> EGCKQIM | T <u>A</u> EGCKQIM | 0.17                        | 0.18                        |
| Total Difference |          |                    |                    | 0.17                        | 0.18                        |

## Clustering of Data

The calculation of immunological distances can be used to compare two sequences or an entire set of sequences. Clustering procedures allow for grouping immunologically similar sequences while at the same time separating more distant ones. Agglomerative clustering is used to progressively assign sequences and groups of sequences that are immunologically similar. Clustering procedures involve the generation of a distance matrix and then performing clustering procedures from the

matrix. The Unweighted Pair Group Method with Arithmetic Mean was used *via* the `hclust` function of the `R` language to construct hierarchical trees. UPGMA is one of the simplest methods to perform hierarchical clustering. The method works by progressively joining sequences and groups of sequences based on the average distance between them. The distance between two sequences can be read from the distance matrix itself, however to measure the distance between two groups of sequences, Equation 2.14 is applied.

$$\frac{1}{|\mathcal{A}| \cdot |\mathcal{B}|} = \sum_{x \in \mathcal{A}} \sum_{y \in \mathcal{B}} d(x, y) \quad (2.14)$$

Where  $\mathcal{A}$  and  $\mathcal{B}$  are sets of sequences,  $|\mathcal{A}| \cdot |\mathcal{B}|$  is the product of the sizes of the sequence sets,  $x$  is a sequence in  $\mathcal{A}$ ,  $y$  is a sequence in  $\mathcal{B}$ . This Equation calculates the average of the sum of the distances between all the elements in both sets in a non-redundant fashion so that when the distance from  $x$  to  $y$  is measured, the procedure won't be repeated for the distance from  $y$  to  $x$ . An example of UPGMA is shown in Figure 2.10 on the next page.

As an example, 20 sequences of HIV protein p17 was mapped for HLA A0201 based epitopes and a comparison matrix constructed. It is generally assumed that the distances are symmetrical, but here the procedure needs to be modified since the symmetry assumption is not met. Thus, the following procedure is followed:

1. Calculate immunological distances between sequences
2. Populate an  $n \times n$  matrix, where  $n$  is the amount of sequences with  $D(i, j)$ , where  $D$  is the distance from sequence  $i$  to  $j$
3. Separately group the rows and the columns
4. Visualise in a heatmap with dendrograms indicating the hierarchical nature of clustering rows and columns

The clustering procedure of 30 sequences of HIV p17 based on predicted epitopes of HLA A0201 will be used to illustrate the process.

### Construction of Dissimilarity Matrices

The calculation of immunological distances is performed between sequences in an all versus all fashion, meaning that if a hundred sequences are to be compared to each other, for each sequence there will be ninety-nine comparisons with all the other sequences. The immunological distance of a particular sequence to itself is calculated as zero. The results of these comparisons are used to populate an  $n \times n$  matrix, where  $n$  is the amount of sequences compared. The term  $D(i, j)$  denotes the distance from sequence  $i$  to  $j$  or, by what immunological degree  $i$  needs to change to

|   |   |   |          |   |   |
|---|---|---|----------|---|---|
|   | A | B | C        | D | E |
| A |   |   |          |   |   |
| B | 4 |   |          |   |   |
| C | 6 | 3 |          |   |   |
| D | 5 | 9 | <b>2</b> |   |   |
| E | 7 | 6 | 5        | 9 |   |

|   |   |
|---|---|
|   | CD                                      |
| A | $\frac{AC+AD}{2} = \frac{6+5}{2} = 5.5$ |
| B | $\frac{BC+BD}{2} = \frac{3+9}{2} = 6.0$ |
| E | $\frac{EC+ED}{2} = \frac{5+9}{2} = 7$   |

(a) UPGMA Step 1

|    |     |     |    |   |
|----|-----|-----|----|---|
|    | A   | B   | CD | E |
| A  |     |     |    |   |
| B  | 4   |     |    |   |
| CD | 5.5 | 6.0 |    |   |
| E  | 7   | 6   | 7  |   |

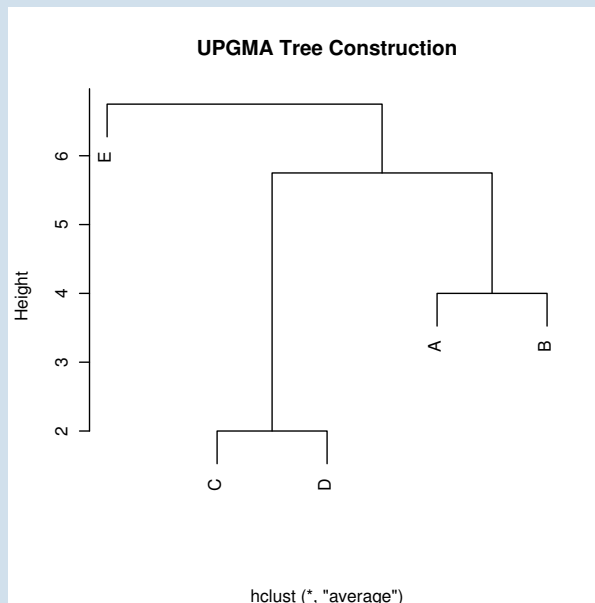
|    |  |
|----|--|
|    | AB   |
| CD | $\frac{CD.A+CD.B}{2} = \frac{5.5+6}{2} = 5.75$ |
| E  | $\frac{E.A+E.B}{2} = \frac{7+6}{2} = 6.5$      |

(b) UPGMA Step 2

|    |             |    |   |
|----|-------------|----|---|
|    | AB          | CD | E |
| AB |             |    |   |
| CD | <b>5.75</b> |    |   |
| E  | 6.5         | 7  |   |

|   |  |
|---|--|
|   | ABCD   |
| E | $\frac{E.AB+E.CD}{2} = \frac{6.5+7}{2} = 6.75$ |

(c) UPGMA Step 3



(d) UPGMA Tree

Figure 2.10: Example of UPGMA process. Five hypothetical sequences with arbitrary distances between them are clustered. In (a) the distance matrix is populated with the distances between the sequences. The closest distance C to D is marked bold and C and D are joined. The new sequence group (CD) is compared to the rest of the sequences via an Equation similar to 2.14 as shown in the right hand panel. A new matrix is constructed as shown in (b). The procedure is repeated with A and B forming the new group AB. The final grouping is shown in (c) to be AB.CD with the final distance of E measured to AB.CD. In (d) the constructed dendrogram from the procedure is shown. The height shows the distances of the sequence groups to their neighbours, e.g the distance between A and B is 4, while the distance between AB and CD is 5.75.



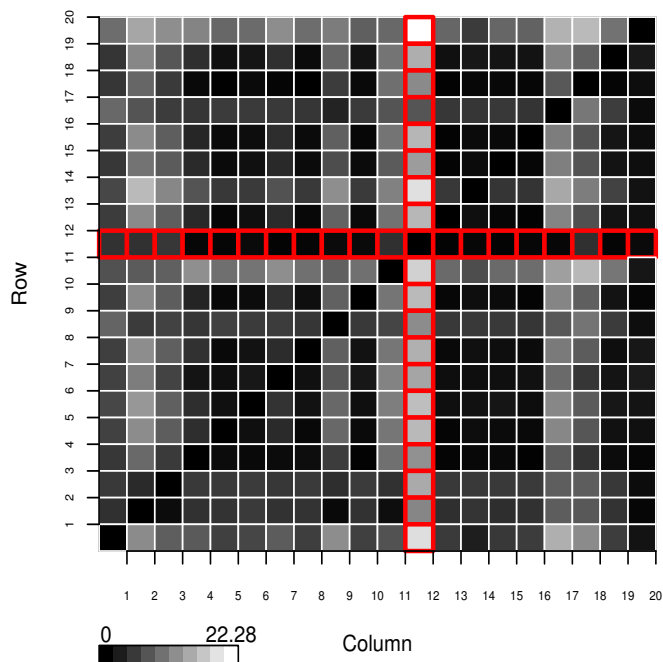


Figure 2.11: Twenty sequences of HIV protein p17 were compared for immunological distances based on potential epitopes related to HLA A0201. This color map is the visual representation of these distance values. The legend indicates the range of the distances. The highlighted portion show the distances for sequence 12, values in the row indicating how far the sequence is to other sequences and the column showing how far the other sequences are to sequence 12. Note the lack of similarity between the row and column distances.

be transformed into  $j$ . If  $i$  were to denote a row in the matrix, it means that all the values in row  $i$  taken as a whole, gives a general pattern of how far sequence  $i$  is to all the other sequences. The matrix can be visualised as in Figure 2.11. Close inspection reveals the general asymmetry of the matrix.

### Effects of Asymmetrical Measurements

With most clustering algorithms, it is assumed that the distance between two objects that are compared, is symmetrical, i.e. the distance  $D(i, j) = D(j, i)$ . As stated, the distance measurements calculated here are generally asymmetrical because of the inherent asymmetry of the Frankild score and the way in which pathway weights are implemented. The symmetry assumed by most agglomerative clustering methods simply means that the rows (or columns) in the dissimilarity matrix are clustered together, for instance, the matrix in Figure 2.10 on the preceding page is symmetrical. Transposing the matrix so that, essentially, the columns are clustered together would yield the same dendrogram as in Figure 2.10 on the previous page(d). To cluster rows, a distance between each of the rows need to be calculated in an all versus all fashion. Here, the euclidean distance metric is used and its calculation shown in Equation 2.15.

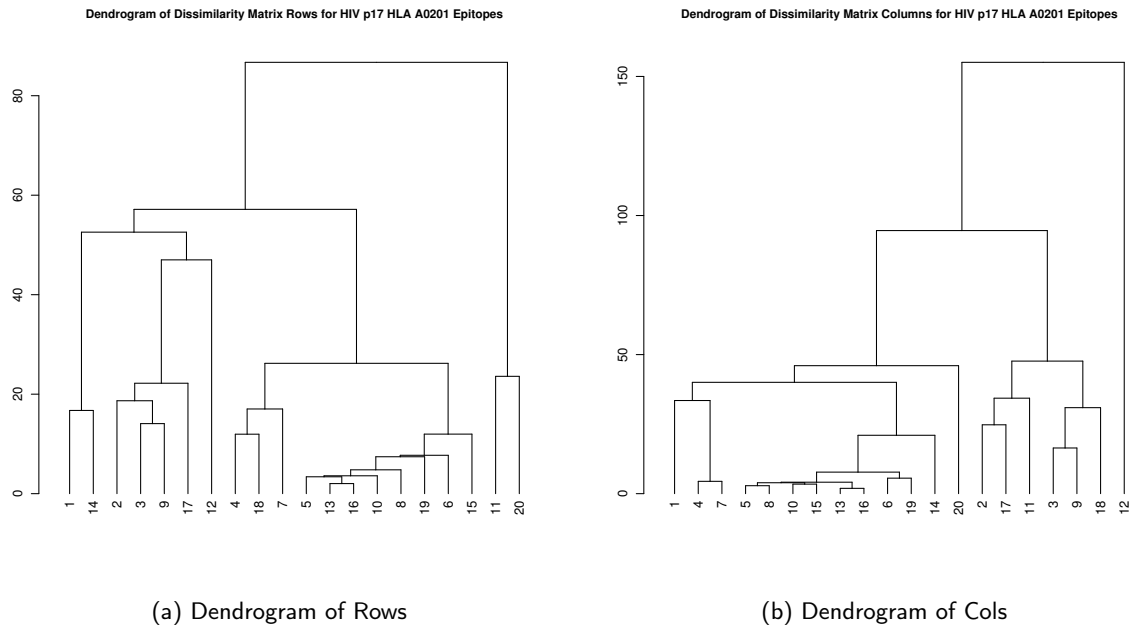


Figure 2.12: The dendrograms were constructed from the distance matrix illustrated in Figure 2.11 on the previous page. In (a) the rows are clustered, grouping together sequences that share similar immunological distances *to* the other sequence. In (b) the columns are clustered together, grouping together sequences that share similar immunological distances *from* other sequences.

$$D_{pq}(p, q) = \sqrt{\sum_{i=1}^n (p_i - q_i)^2} \quad (2.15)$$

Where  $D_{pq}$  is the euclidean distance between row  $p$  and row  $q$ ,  $i$  is the sequence number and  $n$  is the total amount of sequences. A matrix is populated with the results with  $p$  and  $q$  providing the location within the matrix. The newly constructed matrix is symmetric and the UPGMA procedure can be applied to perform the clustering. This procedure is repeated for the columns and an example is shown in Figure 2.12. Alternatively, the maximum distance between two sequences can be used as a single measurement. This results in a symmetric heatmap and therefore the grouping of the rows and the columns would be more equivalent.

## Visualising The Clustering

### Heat Maps

The dendrograms shown in Figure 2.12 is useful for grouping together sequences that are similarly distant from and to other sequences. To better visualise the grouping, where both the row and column clustering are simultaneously visible, a heatmap is used. The colourmap shown in Figure 2.11 on the previous page is an example of a heatmap, but the term heatmap will be used here as the colourmap shown with optimal row and column ordering as determined by hierarchical clustering. An example is shown in Figure 2.13 on page 58.

To determine the optimal amount of groups in which the row and columns can be divided, the

Hubert gamma value is used. The Hubert gamma value is a general measure that, when cutting a dendrogram into a certain amount of groups, how well the groups are separated and how little variance there is within groups (Dunn, 1973). The value ranges between 0 and 1, with higher values indicating better clustering. By cutting the dendrograms stepwise into smaller groups (i.e. larger amount of groups) and measuring the Dunn index at each step, the optimal amount of groups can be determined. The heatmap is annotated by indicating where the partition is between the calculated groups for both the columns and the rows. On the heatmap itself, clustering sectors are formed, making it easier for the researcher to choose sequences that group well together or conversely, sequences that are far apart.

In Figure 2.15 on page 60 the grouping procedure is applied to the matrix. From the heatmap, areas of mass similarity or dissimilarity are easily visible. For instance, the columns that represent sequences 2, 17 and 11 are clustered with the rows representing sequences 15, 6, 19, 8, 10, 16, 13, 5, 7, 18 and 4. This is a *distant* clustering, meaning that the indicated row sequences are very similarly far from the sequences represented by the indicated columns. For the row that represents sequence 12, it can be seen that it is generally close to the other sequences, but the converse is not true and, in fact, in the column dendrogram, it is placed in its own group.

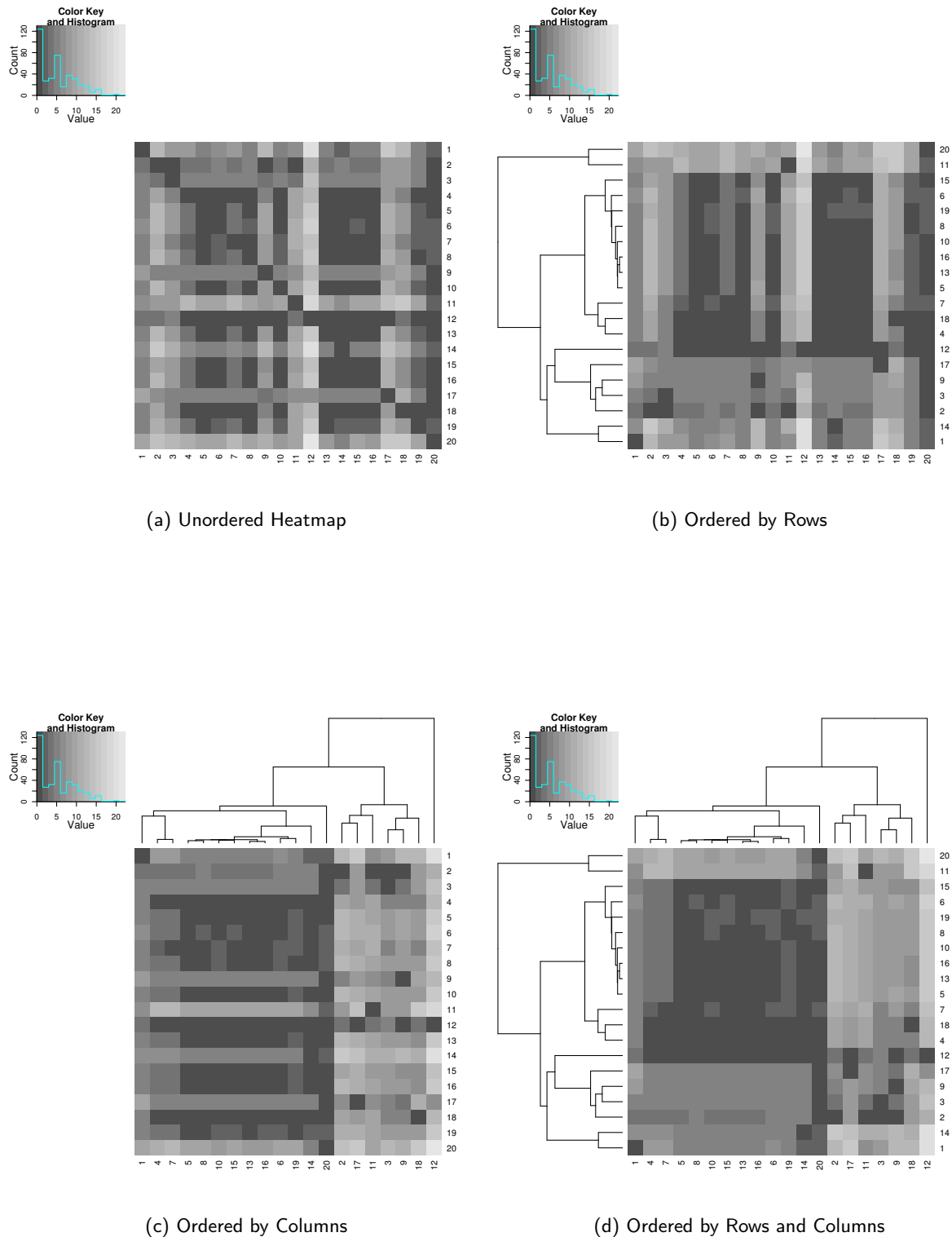
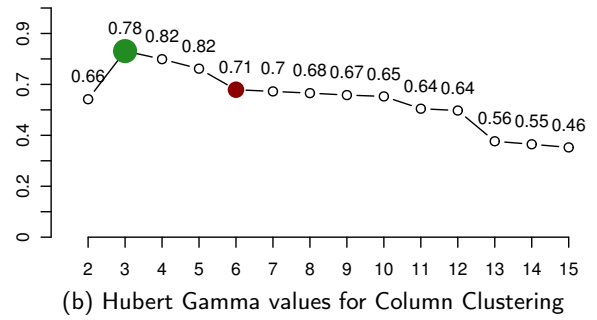
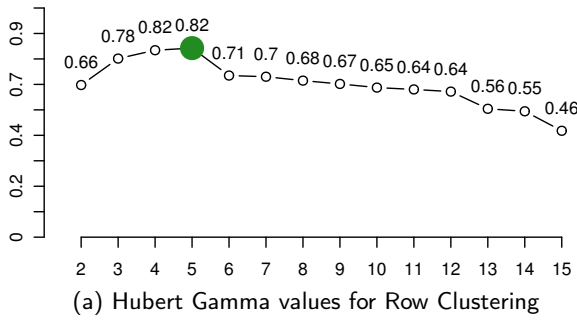
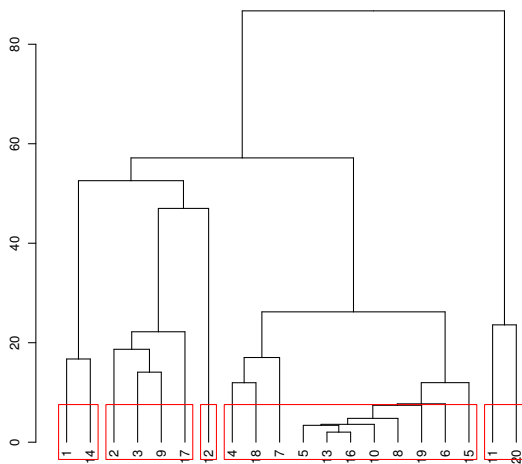


Figure 2.13: These Figures show how the clustering of rows and columns of the distance matrix for the 20 sequences of HIV p17 can be used to reorder the original colour map to clearly indicate sequence grouping. The colour scale is described in the top-left corner of each plot, with darker colours indicating further distances and darker colours, closer distances. The unordered matrix is shown in (a). In (b) and (c) ordering by row and then by column is shown and in Figure (d) heatmap with ordered rows and columns are shown. Note how the general trend in shading matches with the branch lengths of the dendrograms.



Dendrogram of Dissimilarity Matrix Rows for HIV p17 HLA A0201 Epitopes



Dendrogram of Dissimilarity Matrix Columns for HIV p17 HLA A0201 Epitopes

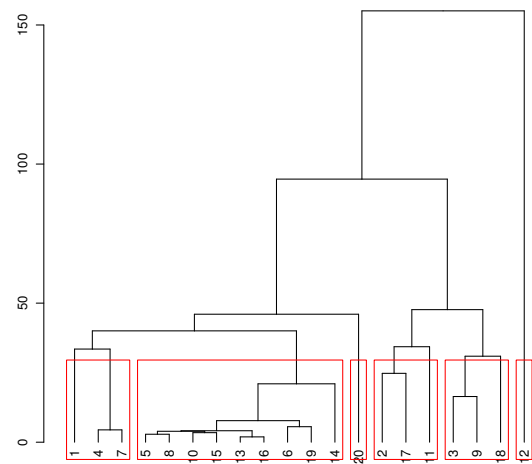


Figure 2.14: Grouping of heatmap clusters. To annotate the heatmap with groupings as shown in Figure 2.15, the Hubert gamma values are calculated according to how many parts the dendrograms are split into. In (a) and (b) this is shown as a plot with the x-axis and y-axis representing group numbers and Hubert gamma value respectively. The predicted optimal number of groups is shown as a green dot. The red dot in (b) indicates that the optimal group number measurement of 3, has been overridden with 6. In (c) and (d) the groupings are shown with the Rows dendrogram cut into 5 parts and the Columns dendrogram cut into six.

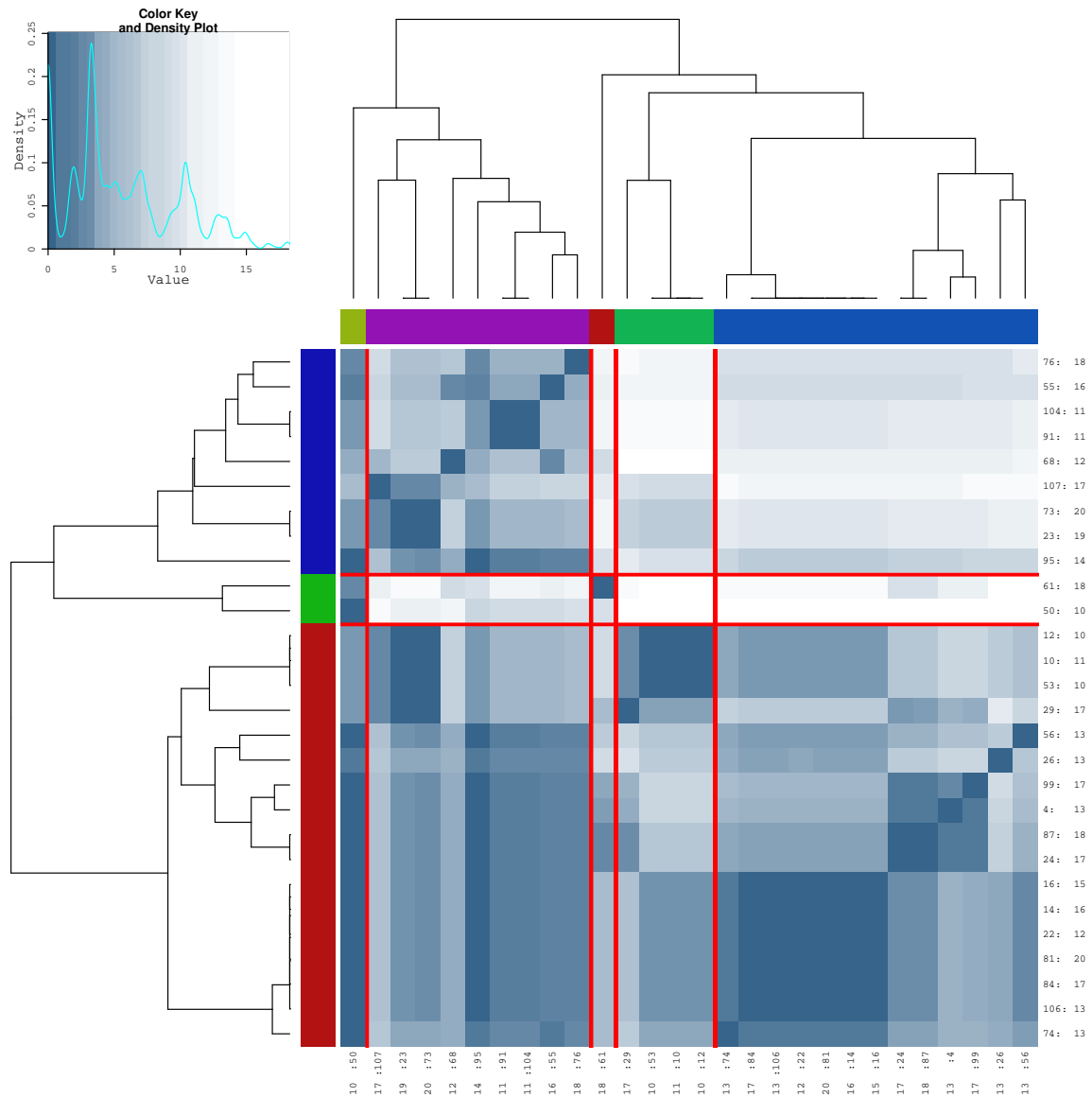


Figure 2.15: Heatmap of certain predicted HLA\*A0201 restricted epitopes with grouping annotation. The heatmap is annotated with colours representing groups of each dendrogram. The heatmap is also 'cut' between the groups to make clustering of Rows and Column samples clearer. The density plot in the top left corner is a visual indication of overall similarity/dissimilarity. The colour scale is calculated according to 'breaks' determined from the original matrix data. A linear colour scale may fail to ephasize regions of exceptional difference or similarity.

**Detailed Immunological Comparison** The difference between groups of sequences be further elucidated by tabulating the epitope differences and similarities between them. Any groups of sequences can be compared to each other, though importantly, sequences that form immunologically similar or distant sectors can be read from the heatmap. This will give a clearer picture of the relationship between the sequences. The procedure is divided into two main parts. First, the sequences are scanned for unique epitope sequences and the existence of these epitopes are mapped between the two sequence sets. The frequencies of these epitopes are calculated as well as a weight that can be any combination of the pathway prediction scores. The next step is to combine all the epitopes that occur at specific positions for specific HLA allotypes. The next step is to summarise the relationships between the sequence sets according to predicted epitopes at specific positions in order to compare the epitope variants between the sequence sets. This information is then tabulated according to HLA allotype and position. The frequency of epitopes occurring at the positions as well as the queried weights for these epitopes are utilised to elucidate differences between the sequence sets. An example is shown in Figure 2.16 on the following page. See the description for information on the data displayed. Figure 2.16a on the next page shows the occurrence of specific predicted epitopes in both sequence sets. It may seem redundant to include the differences in weights of the same predicted epitopes, after all they will have the same MHC IC50 and immunogenicity predictions. The weights used here include all the pathway predictions made. TAP and proteasomal predictions can exist beyond the range of the predicted epitope's sequence and for some of the displayed predicted epitopes, the influence of mutations relevant to proteasomal cleavage and/or TAP affinity is indeed evident for the top half of epitopes in the Table. In Figure 2.16b on the following page the total differences between epitopes occurring at a specific position is shown. The comparison here was between immunologically similar sequences, so the SeqLogos do appear similar. It is revealed, that the sequences are not entirely identical, although the differences in epitopes "repertoire" shown here are trivial.

## 2.4 Conclusion

Here, the main development aspects of Fortuna were described, with special focus on implementation of individual C1APP associated predictors, the design of a novel TAP affinity predictor and combination of prediction results. Additionally, it was shown how the output from the prediction results could be used to perform subsequent analyses relating to Clustering of sequences based on immunological profile, entropy/frequency analysis and self-epitope discovery. The following chapter illustrates the interface to these tools via a web-based interface and usage examples of Fortuna.

| Allotype | Length | Pos | Seq        | IC50 | Imm | Weight (1) | Weight (2) | w1 - w2 | f1   | f2   | f1-f2 | Rel Seq   |
|----------|--------|-----|------------|------|-----|------------|------------|---------|------|------|-------|---|
| A1101    | 10     | 42  | IVVASRELER | 177  | 1.0 | 3.64       | 3.81       | -0.17   | 0.43 | 0.14 | 0.29  | 37 4 65 56 27 42 10 30 9 66 43 64<br>15 43 22                     |
| A1101    | 10     | 26  | KIRLRPGGRK | 250  | 2.0 | 3.78       | 3.88       | -0.10   | 0.04 | 0.05 | -0.01 | 37<br>45  |
| A1101    | 10     | 26  | RIRLRPGGKK | 498  | 2.0 | 3.59       | 3.63       | -0.04   | 0.14 | 0.05 | 0.10  | 70 19 42 64<br>24   |
| A1101    | 10     | 85  | SLFNTVATLY | 84   | 0.0 | 3.24       | 3.27       | -0.03   | 0.46 | 0.41 | 0.06  | 39 37 76 4 6 65 56 27 10 9 28 43 64<br>14 39 60 45 43 28 68 22 31 |
| A1101    | 10     | 85  | SLYNTVATLY | 69   | 0.0 | 3.38       | 3.38       | -0.01   | 0.32 | 0.27 | 0.05  | 70 5 19 42 72 30 38 47 66<br>15 41 57 11 63 24                    |
| A1101    | 9      | 8   | MGARASILK  | 467  | 2.0 | 1.03       | 1.03       | 0.00    | 0.04 | 0.09 | -0.06 | 43<br>43 22   |
| A1101    | 9      | 11  | RASILKGGK  | 132  | 2.0 | 1.58       | 1.58       | 0.00    | 0.04 | 0.09 | -0.06 | 43<br>43 22   |
| A1101    | 9      | 11  | RASILRGEK  | 44   | 2.0 | 2.06       | -Inf       |         | 0.25 | 0.00 | 0.25  | 6 65 70 56 10 72 54   |
| A1101    | 9      | 11  | RASILRGGK  | 64   | 2.0 | 1.90       | 1.90       | 0.00    | 0.21 | 0.59 | -0.38 | 75 19 38 66 21 25<br>14 62 15 41 57 60 45 21 25 26 31 75 24       |
| A1101    | 9      | 11  | RASILSGGK  | 51   | 2.0 | 1.99       | 1.99       | 0.00    | 0.04 | 0.05 | -0.01 | 28<br>28  |

(a) Comparing all epitopes

| Allotype | Position | Length | f1   | f2   | w1   | w2   | w1-w2 | Seq List 1   | Seq List 2   | Seqlogo 1 | Seqlogo 2 |
|----------|----------|--------|------|------|------|------|-------|--|--|-----------|-----------|
| A1101    | 11       | 9      | 0.93 | 0.95 | 1.99 | 1.89 | 0.10  | RASILSGGK (1)<br>RASILRGEK (7)<br>RASILKGGK (2)<br>RASILRGGK (1)<br>RASILRGGK (6)<br>RASILRGEK (8)<br>SASILRGEK (1)<br>RASILKGEK (1) | RASILRGEK (4)<br>RASILKGGK (2)<br>RASILRGGK (13)<br>RASILSGGK (1)<br>RASILKGEK (1)                       |           |           |
| A1101    | 26       | 10     | 1.00 | 0.91 | 2.96 | 2.85 | 0.11  | KISLRPGGKK (1)<br>RIRLRPGGKK (4)<br>KIRLRPGGKK (20)<br>KIKLRPGGKK (1)<br>KIRLRPGGRK (1)<br>RIKLRPGGKK (1)                            | KIRLRPGGRK (1)<br>KIRLRPGGKK (17)<br>RIRLRPGGKK (1)<br>RIKLRPGGKK (1)                                    |           |           |
| A1101    | 25       | 9      | 0.93 | 0.91 | 1.18 | 1.19 | -0.01 | KIRLRPGGK (20)<br>KISLRPGGK (1)<br>RIRLRPGGK (4)<br>KIKLRPGGK (1)  | KIRLRPGGK (18)<br>RIKLRPGGK (1)<br>RIRLRPGGK (1)   |           |           |
| A1101    | 85       | 10     | 0.93 | 0.86 | 3.47 | 3.66 | -0.19 | SLYNTVATLY (9)<br>SLFNTVAVLY (1)<br>SLYNTIATLY (1)<br>SLFNTVATLY (13)<br>SLFNTIATLY (1)<br>SLYNTVAVLY (1)                            | SLYNTVAVLY (1)<br>SLYNTVATLY (6)<br>SLFNTVAVLY (1)<br>SLYNTIATLY (1)<br>SLYNTIATLY (9)<br>SLYNTIATLY (1) |           |           |

(b) Comparing epitopes at specific positions

Figure 2.16: The above Figures depict comparisons between two immunologically related sequences. In (a) a snippet of the total list of epitopes is shown. The headings reveal the HLA allotype length of the epitope, aligned position, sequence, predicted IC50 value for MHC binding, POPI prediction, weight associated with the epitope (in this case, the combined predictions for MHC, Proteasomal, TAP and Immunogenicity), the differences between the weights of sequence set 1 and sequence set 2 ( $w1 - w2$ ), the frequency that the specific epitope occurs in the sequence set and the difference between the frequencies. The last column show the sequence numbers that contain the epitope. In (b) a summary of the epitopes occurring at specific positions for associated HLA molecules is shown. The frequencies and weights are read similarly as in (a), although it should be noted that the *average* weight for all the epitopes in the sequence set of the HLA allotype in question occurring at the indicated position is used. The yellow highlighted epitopes in the “Seq List x” columns are epitopes that occur in both sequence sets. The final two columns are the SeqLogos of the epitopes predicted for the two sequence sets.



## Implementation of Fortuna

In this chapter, it will be demonstrated how the aforementioned prediction and analysis tools can be made accessible to the general user. The design, method of operation, visualizations and interfaces will be discussed as well as examples shown. First, the implementation of Fortuna as a web-based tool will be discussed by describing the server-side and client-side processes. This is followed by a usage example of Fortuna.

### 3.1 Fortuna as a Web-Based Application

In order to make full use of predictions and analysis tools provided by created, a suitable interface to them is needed. Countless problems can be encountered when developing a stand-alone package that provides interface to these tools. Problems include, but not limited to, computer architecture, operating system and available computer resources. It is for this reason that a web-based application is designed with a server backend. Any person can access the tool, provided they have a compatible web browser and internet connection. The tool, as a whole, is called *Fortuna*. It is divided into two parts:

1. *Client Side* - The web browser from which the tool is accessed. Sequence input, parameters and analyses are queried from here.
2. *Server Side* - Where the predictions, analyses and outputs are generated and sent to the browser.

Through constant interface between the Server and Client side, analyses are performed and output generated. A flow diagram of the Client/Server interface is shown in Figure 3.1 on the next page.

The design, however, is not a trivial process and a multitude of programming packages are employed to accomplish the task. First, we will examine the input and parameters required for initial analysis, i.e. immunological predictions. Next, we will look at the programming tools

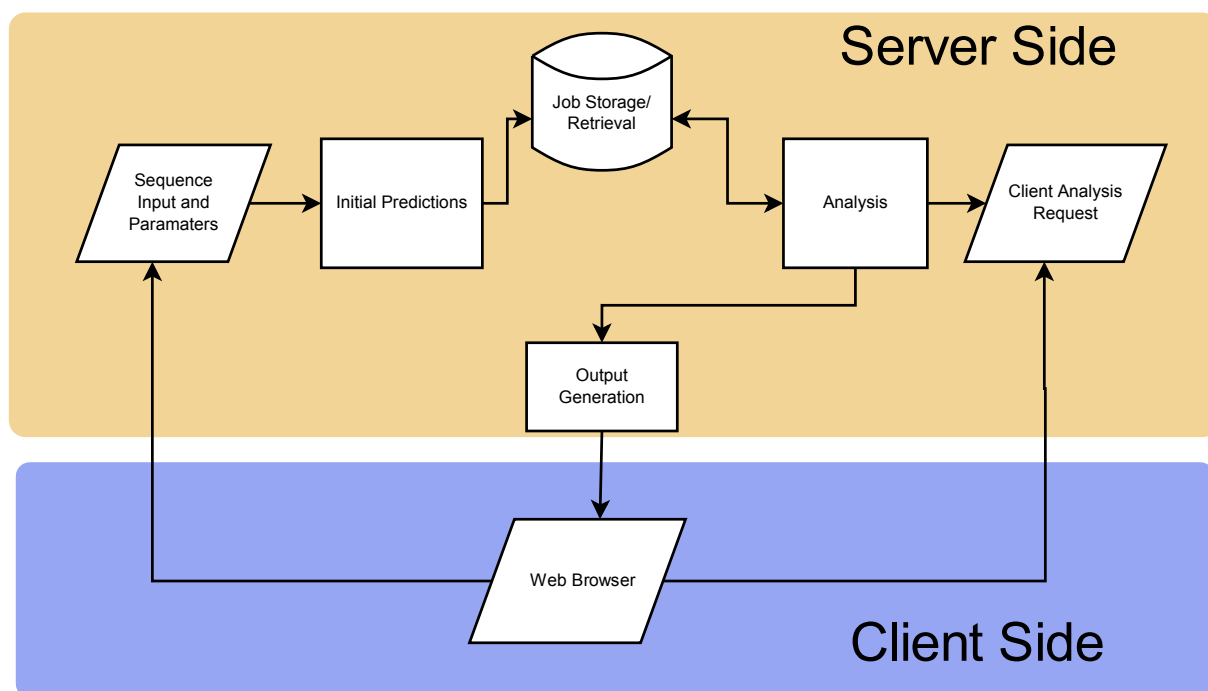


Figure 3.1: Flow diagramme of Fortuna Client/Server interface. All requests are sent *via* user from the web browser to the server. For initial predictions, where the user sends sequences to be analyzed according to a set of parameters, the server processes these requests and stores the submitted job. Upon completion, the user can access a job and request different analyses to be performed on it. The output is then sent to the browser.

utilised to accomplish the internal programming that include interface to predictions and analysis as well as creation of the server side that will handle requests from the user. Lastly, the design and use of the interface will be discussed.

### 3.1.1 Server Side Processes

#### How Predictions are Handled

Immunological predictions in a single sequence is a relatively simplistic task, but since the intent here is to reveal immunological information across multiple sequences, additional steps need to be taken to ensure predictions are done in a consistent manner. The overall steps and applications/packages associated with each are listed in Table 3.1 on the following page. This is especially true if the sequences differ in length. The solution is to first align the sequences with `clustalw2` and map the positions in the native sequences to their ‘aligned’ positions. So, if a predicted MHC ligand, proteasomal cleavage site, etc. occurs at position *X* in a sequence, this position is then mapped to its position within the alignment. All the positional references between the sequences thus remain consistent.

All of the predictions made rely on words of amino acids as input. The word list consisting of different length words are created. A summary of the word lengths and steps associated with them is shown in Table 3.2 on page 66. Proteasomal Cleavage predictions always use words of length 10, while words MHC- and TAP affinity predictions can be variable. The word list for

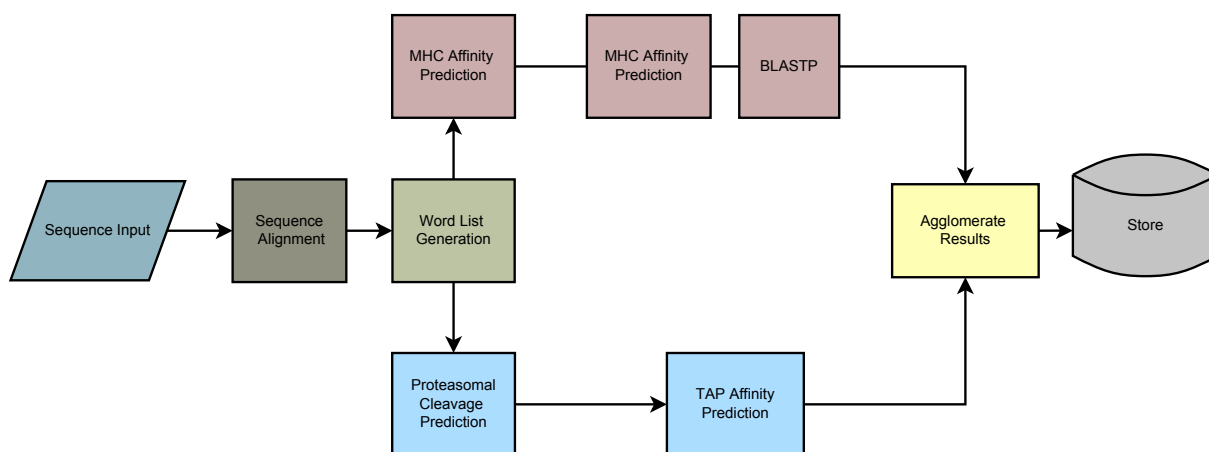


Figure 3.2: Overview of the prediction process. Sequences and parameters (HLA allotypes, ligand lengths, thresholds etc.) are taken as input. The sequences are validated and those that pass the validation test are aligned using `clustalw2`. Word lists are generated and sent to the respective prediction tools. Proteasomal cleavage prediction using 10-mer words, while the length of words used for NetMHC is determined by the input parameters. The results are agglomerated and stored.

Table 3.1: Steps in Prediction Process.

| Step | Process                         | Application/Package Used   |
|------|---------------------------------|--|
| 1    | Sequence Alignment              | The <code>clustalw2</code> package <i>via</i> the BioPython library                                    |
| 2    | Word list generation            |  |
| 3    | MHC Affinity Prediction         | The stand-alone version of NetMHC-3.0 <i>via</i> a python interface                                    |
| 4    | Proteasomal Cleavage Prediction | The SMM application using proteasomal cleavage matrices  |
| 5    | TAP Affinity                    | VLTAPP implemented in the R statistical language and accessed <i>via</i> the <code>rpy2</code> package |
| 6    | Immunogenicity                  | An R implementation of the POPI tool   |
| 7    | Self-Epitope Prediction         | The NCBI's BLAST+ package utilising a 2009 version of the RefSeq database                              |

TAP affinity prediction is dependent on the proteasomal cleavage prediction steps.

After all the prediction steps have completed, the results are agglomerated and stored on disk for future retrieval by the user.

### Management of Jobs and Storage

Fortuna is not intended to be a fully open system. This means that a user needs to register and that jobs submitted by a particular user are not accessible by other users to ensure a certain level of privacy. This also means that jobs that are submitted by a user can be retrieved at any other time. The completed jobs are stored in a *pickled* format. In the Python programming language, objects are entities of variables that can contain potentially any computer-encodable information. All the prediction results are stored in Python objects. The `Pickle` library for Python can be used to store any object on disk. Each submitted job has a unique identification number associated

Table 3.2: Word List Generation

| Step                            | Word Length   |
|---------------------------------|---|
| Proteasomal Cleavage Prediction | 10-mer  |
| TAP Affinity                    | Ranging from 9-mer to typically 20-mer  |
| MHC Binding Affinity            | NetMHC allows for binding affinity prediction of 8-11 mer words                           |
| POPI                            | Can be arbitrary length, though the same word list is used as for MHC affinity prediction |
| BLAST                           | Same length as for MHC affinity prediction  |

Table 3.3: Programming Package Utilised for Server Side Development.

| Task Type                                  | Tool/Interface                       | Citation   |
|--|--------------------------------------|--|
| Sequence Alignment and FASTA file handling | clustalw2<br>BioPython               | <i>via</i><br>(Thompson <i>et al.</i> , 1994, Cock <i>et al.</i> , 2009) |
| Cluster analysis                           | R-packages fpc and stats             | (Hennig, 2009, R Development Core Team, 2009)                            |
| Graphical Output                           | R-packages gplots, plotrix and Cairo | (Warnes, 2009, Jim Lemon <i>et al.</i> , 2009, Urbanek and Horner, 2009) |
| VLAPP ANN                                  | AMORE                                | (Limas <i>et al.</i> , 2007)   |
| Web Application Framework                  | Turbogears                           | (Ramm <i>et al.</i> , 2006)  |

with it. Referenced from the identification number is the filename that contains the ‘pickled’ Python object. Only jobs that have been complete are accessible for analysis.

### Server Side Development

The Python version 2.6 and R version 2.9.1 programming languages were mainly used for server side development. The Turbogears package [<http://www.turbogears.com>] was used to create the Web server. Turbogears allows for rapid development of web applications and has intrinsic methods for dealing with databases as well as web browser-compatible content generation. Most of the analysis is done in Python using mostly novel code. The handling of FASTA files as well as Clustalw2 alignment was accomplished through the BioPython, though in the latter case the code needed to be modified. Graphical output for the Heatmaps, Treemaps, Entropy/Frequency graphs and Hubert gamma values are generated with R. The python interface to the WebLogo application is used for creation of the SeqLogos. VLAPP uses an ANN implementation provided by the AMORE package (Limas *et al.*, 2007).

### 3.1.2 Client Side

Although the output is generated on the server side, some browser-based code need to be executed to provide an appropriate layout of Figures, Tables etc. Web browsers use **JavaScript** as a programming language to manipulate the documents they display. The **JQuery** package [<http://www.jquery.com>] was used as the JavaScript framework to accomplish most of the JavaScripts routines needed for the client interface. Layout and aesthetics are handled by the **JQueryUI** framework [<http://www.jqueryui.com>]. This library provides tools to design an easier accessible Graphical User Interface (GUI) on the browser side, which is traditionally a difficult task. The **DataTables** packages [<http://www.datatables.net>] was used to enhance the Tables that display output from various analyses. This package not only optimally render the Tables (that can be very large), but also provided filtering routines if the user decided to look for specific entries. For instance, when the epitope prediction output is generated, the user might want to view a specific epitope and can filter the rows according to that epitope's sequence. The **DataTables** package also allow for the saving of tabular data, making further use of the predictions outside the context of *Fortuna* a lot easier. Indeed, the data in Table 2.11 on page 50 was obtained by directly saving the data from the "Epitope Info" analysis Table in the browser.

In order not to intrude on the user while performing different analyses, the generation of dynamic content (i.e. output) is handled *via* AJAX requests. AJAX allows the web browser to communicate with the server in a 'behind the scenes' fashion allowing multiple analyses to be performed simultaneously and avoiding content-loss that is associated with refreshing or changing the page of the web browser.

## 3.2 Example of Interface and use of Fortuna

To illustrate the interface and use of *Fortuna*, the process of submitting the sequences an parameters to examples of analyses will be shown. However, to avoid redundancy, not every conceivable analysis will be performed. The demonstration will be limited to:

1. Registration
2. Starting a new job
3. Job overview
4. Clustering analysis
5. Comparing sets of sequences

This will only show a fraction of the use of *Fortuna*, but later in the text when analyses will be performed on HIV and Influenza, the use of all the tools will be discussed in more detail. The

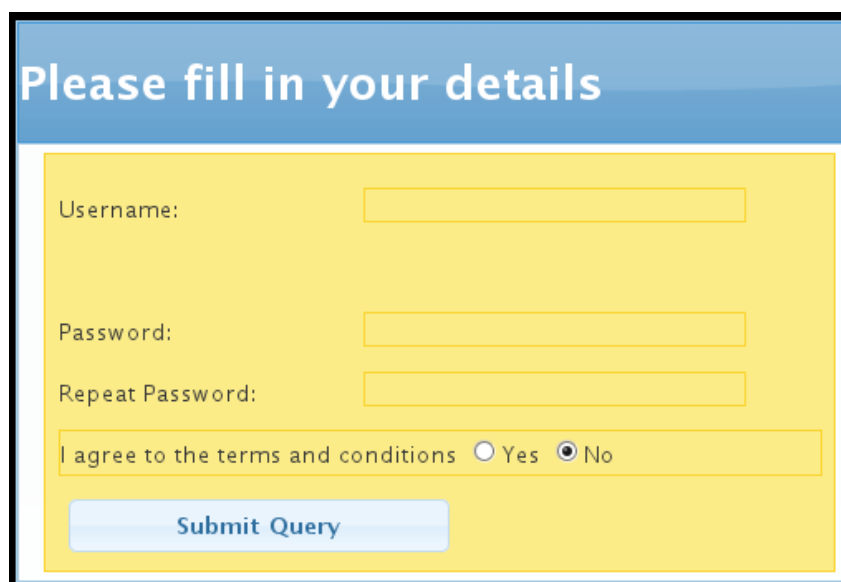
The image shows a registration form titled "Please fill in your details". It has a blue header with the title in white. The form area has a yellow background. It contains four input fields: "Username:", "Password:", "Repeat Password:", and a checkbox area for "I agree to the terms and conditions" with radio buttons for "Yes" and "No". The "No" option is selected. Below the form is a blue button labeled "Submit Query".

Figure 3.3: Registration Screen

examples that follow only demonstrate the use and to a certain degree, the technologies used to make *Fortuna* possible.

### 3.2.1 Registration

To submit a job and access any of the tools, a user needs to register. This ensures that previously submitted jobs can be retrieved in an organised manner, i.e. not jumbled with other users' jobs. A more important point, is that it ensures a reasonable level of privacy. Only the owner of a job can access it at any given time. *Fortuna* uses NetMHC under an academic license, meaning that the user has to agree to certain terms and conditions in order to register, most notably that *Fortuna* may not be used for commercial purposes. The registration screen is shown in Figure 3.3.

### 3.2.2 Starting a New Job

A new job is started at the screen shown in Figure 3.4 on the following page. Each user can only view his or her jobs and this provides a level of organisation as well as privacy. It should be noted that some of the options like the HLA allotypes and lengths of the MHC ligands are set parameters and cannot be changed after the job has been submitted. After submitting the sequences, the user is shown a confirmation dialog, giving information on the validity of the uploaded sequences. Non-standard amino-acid or arbitrary characters within the sequence portion of a FASTA entry are not tolerated and these sequences are automatically omitted from predictions. The files are also tested if they are, indeed, valid FASTA files. If this condition is not met or if all the sequences within the file contains 'invalid' characters, the job cannot be started.

Completed Jobs | **New Job** | Settings

Enter a job title (default: time and date of job submission)

Select allotype(s): A0101, A0201, A0202, A0203, A0204, A0206, A0211, A0212, A0216, A0219

Select Length(s): 8, 9, 10, 11

Input sequence(s) in a Fasta format.

Fasta File: Browse...

MHC binding threshold IC50 nM (optimal = 500 nM): 500

Proteasome threshold as fraction (usual = 0.05): 0.27

Max length of proteasomal fragments (typical = 20-30): 20

Submit Query

(a) Defining a new Job

|                          |                     |   |    |
|--------------------------|---------------------|---|----|
| <input type="checkbox"/> | 2010-06-24 23:41:46 | pol_a02_a30_a68_a02_b58_b15                   | 91 |
| <input type="checkbox"/> | 2010-06-24 19:56:37 | gag_a02_a30_a68_a02_b58_b15                   | 90 |
| <input type="checkbox"/> | 2010-06-22 15:44:24 | [hiv_p17_p24_integrase_gp160_protease_vpr_zs] | 70 |
| <input type="checkbox"/> | 2010-06-20 13:06:37 | integrase test                                | 65 |
| <input type="checkbox"/> | 2010-06-17 20:48:49 | flu europe joined                             | 64 |
| <input type="checkbox"/> | 2010-06-11 20:21:25 | FLU Europe pb1                                | 63 |
| <input type="checkbox"/> | 2010-06-10 21:54:48 | flu europe np (little)                        | 62 |
| <input type="checkbox"/> | 2010-06-10 21:24:06 | flu europe np trim                            | 61 |
| <input type="checkbox"/> | 2010-05-18 20:48:42 | Tue 18 of May 2010 at 20:45                   | 60 |
| <input type="checkbox"/> | 2010-05-06 17:33:24 | HIV ZA p17 for HLA A*02                       | 59 |
| <input type="checkbox"/> | 2010-04-28 17:31:52 | Wed 28 of April 2010 at 17:30                 | 58 |
| <input type="checkbox"/> | 2010-04-07 15:02:56 | HIV p17 ZA                                    | 57 |
| <input type="checkbox"/> | 2010-04-07 14:00:38 | Wed 07 of April 2010 at 13:59                 | 56 |

(b) The completed jobs in a list

Figure 3.4: In (a) a new job is defined by uploading or entering sequences defined and setting HLA, Proteasomal and TAP parameters. Different predictable allotypes can be selected as well as the chosen MHC ligand length(s). The default proteasomal threshold is also set here as well as the maximum size of the TAP lengths. After the parameters are set, the job can be submitted. In (b) a list of completed jobs are shown.



Figure 3.5: The overview of the completed job shows what HLA allotype ligands have been predicted as well as the lengths concerned. Other parameters like the Proteasomal threshold value and maximum TAP ligand length is shown. The aligned sequences are also displayed. At the top of the image, the various tabs can be seen that are used to access the analysis tools. The content is dynamically shown, meaning that setting the parameters on a specific page will not be changed if navigating to other analysis tools.

### 3.2.3 Overview of the Submitted Job

After the completion of the alignment, predictions and BLAST analysis of the input job, it can be viewed as shown in Figure 3.5. The overview mainly shows the parameters used for prediction analyses as well as the predictions made for the sequences. With modified code of the Sequence Manipulation Suite version 2, the aligned version of the sequences are also displayed (Stothard, 2000). From this window, all the analysis tools can be accessed.

### 3.2.4 Example of Cluster Analysis

One of the design goals of Fortuna is to make the analysis tools as flexible as possible. In the extreme case, defining the options of clustering is a good example of this. In Figure 3.6 on page 72 the process of selecting clustering options and then the output produced is shown. As described, multiple sets of options can be passed to the server for analysis. The sets of parameters are handled separately when clustering analysis is performed on the server. For each parameter set, a distance matrix is produced and these matrices are added together to produce the final distance matrix. The rationale behind the option to separate the parameters is mainly due to



accuracy differences in accuracy between neural networks for their associated HLA allotypes. MHC ligand predictors tend to have different optimal IC50 cutoff values (though biologically, the viable IC50 value remains virtually the same). The user thus has the option to give different HLA predictions different threshold values if epitopes of different HLA allotypes are analyzed simultaneously.

The output consists of a Heatmap, banner plot, dendrograms and grouping information. The Heatmap shows the general clustering pattern of the sequences. The banner plot is a visual representation of the clustering and essentially show how appropriate agglomerative clustering is for immunological distances between the sequences. The dendrograms show how the rows and columns of the heatmap cluster together and exist merely as a reference. The grouping information tab shows the Dunn-indexes for different group numbers of both the rows and the columns.

For further analysis of the difference between sets of sequences, the “Compare Sequences” tool is used. Sets of sequences that can be seen to be immunologically close or distant on the heatmap can be selected from the side columns and the sequence numbers copied to the “Compare Sequences” tool, which will be discussed next.

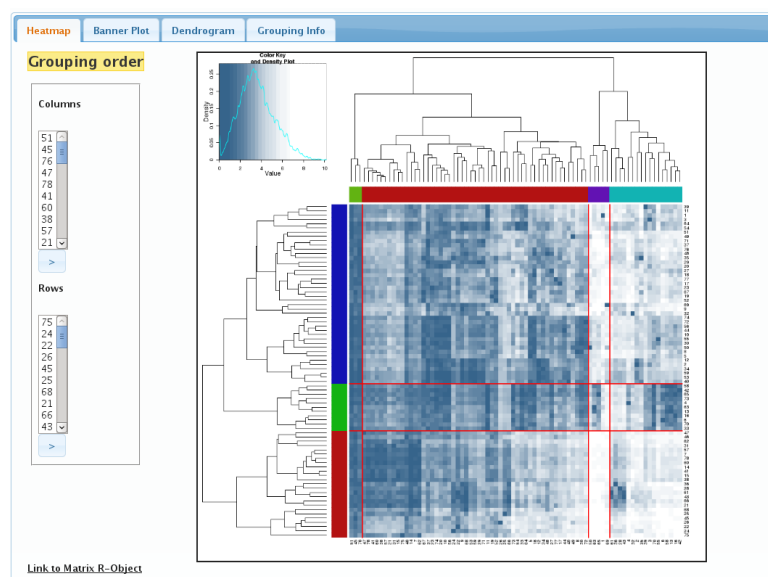
### 3.2.5 Comparing Sequences

Any sequence or set of sequences can be compared based on their CTL epitope profile. The interface and part of the output is shown in Figure 3.7 on page 73. What should be noted from this figure, is the HLA allotype, ligand length as well as the thresholds for MHC affinity, proteasomal cleavage and TAP affinity thresholds. These parameters can be tweaked in all the analyses. The user may, in conjunction to the mandatory MHC affinity threshold, choose other steps in the pathway deemed relevant to the fate of the potential MHC ligand. A ligand with appropriate affinity may in its original protein never be cut at the correct C-terminal end or be transported via TAP to the ER. In the context of the “Compare Sequences” analysis, the parameters can be set according to the parameters used in clustering. Note, however, that only one parameter set can be compared at a time and that this analysis is not intended to compare sequences that have been clustered based on different thresholds and weights. Still, simultaneous comparison of multiple HLA allotype ligands can be performed. The user also chooses the weights to be used comparing the sequences together. The last option is the sets of sequences to be compared. The output is obtained and displayed in the bottom half of the analysis screen where the user can examine the existence of particular potential epitopes in the different sequence sets or, alternatively, examine the sets of different epitopes that occur for each allotype at a particular position. A more detailed explanation of the output has been explained

The screenshot displays the Fortuna web interface. The top section, titled "Remove Options", contains a table with columns: Allotypes, Lengths, Used Cut-offs, and Weights. The "Weights" column is further divided into M, Prot., and TAP, with sub-columns for (IC50) Type Value and Immunogenicity Total. Two rows are visible: A1101 and A0201. The A0201 row is selected with a checkmark. Below the table, there are input fields for "Groups (R/C): 3,2", "Missing Epitope Score: 0.3", and "Weighted: Max Difference". "Submit" and "Update" buttons are at the bottom.

The bottom section, titled "Sequence Selection", features a table with columns "Select", "Deselect", and "Title". A search bar is present. The table lists seven sequences with their identifiers and amino acid sequences. The first sequence is selected with a checkmark.

(a) Setting Clustering Options



(b) Clustering Output

Figure 3.6: In (a) the clustering parameters are set. Different scores from the pathway can be selected to serve as weights for predicted epitopes. Thresholds can also be set. It is important to notice that multiple sets of parameters can be passed. In this example, two HLA allotypes are chosen individually and added to the total options. The bottom half of (a) shows the sequence list. This table allows the user to select desired sequences for cluster analysis as well as providing filtering options if a specific set of sequences are required. In (b) the output is shown. The results are also arranged in tabs, with the main display concerning the heatmap, and the rest of the tabs dealing with visual representation of clustering quality and grouping information. The numbers to the left of the heatmap are the index numbers of the sequences as arranged in the rows/columns of the cluster. Selected sequences can be copied to another analysis tool, "Compare Sequences".

The screenshot displays the Fortuna web interface for comparing sequences. The top navigation bar includes tabs for Overview, MHC TreeMap, Epitope Info, Compare Sequences (active), Cluster Analysis, Entropy/Frequency, and Self-Epitopes. The main interface is divided into two sections: a query input area and a results table.

**Query Input Section:**

- Allotype Selection:** A list of allotypes (A0216, A0219, A0301, A1101, A2301) with a dropdown menu set to A1101.
- MHC:** A slider set to 500.
- TAP Origin:** A dropdown menu set to None.
- Profession:** A dropdown menu set to None.
- Immunogenicity:** A dropdown menu set to Immunogenicity.
- Sequence Lists:** Two lists of amino acid positions and their corresponding sequences.
 

| Sequence List 1            | Sequence List 2              |
|----------------------------|------------------------------|
| 75 24 22 26 45 25 68 21 66 | 4 32 2 35 39 3 70 55 6 58 13 |
| 43 61 28 36 38 15 41 14 60 | 16 42                        |
| 78 7 57 31 62 46 47        |                              |

**Results Table:**

The results table shows the following data:

| Allotype | Length | Pos | Seq       | IC50 | Imm | f1   | f2   | f1-f2 | Rel Seq |
|----------|--------|-----|-----------|------|-----|------|------|-------|---------|
| A0201    | 9      | 34  | KQYHLKHLV | 148  | 0.0 | 0.04 | 0.00 | 0.04  | 7       |
| A0201    | 9      | 34  | KTYHDKHIV | 324  | 2.0 | 0.04 | 0.00 | 0.04  | 24      |
| A0201    | 9      | 34  | KTYHLKHLV | 198  | 0.0 | 0.04 | 0.00 | 0.04  | 31      |
| A0201    | 9      | 36  | YHIKIVWA  | 6    | 0.0 | 0.04 | 0.00 | 0.04  | 24      |

Figure 3.7: Comparing Sequences example

in Figure 2.16 on page 62.

### 3.3 Conclusion

In this chapter the design and implementation of the Web-accessible tool, *Fortuna* was discussed. It was illustrated how freely available software packages can be used to create sophisticated interfaces to access the prediction, analysis and visualization methods used by *Fortuna*. In the following chapter, it will be shown how *Fortuna* can be used to perform real-world prediction analysis of CTL epitopes, using Influenza A and HIV as pathogens for investigation.

The main aim of this project is to provide an accessible tool to aid in the study of potential CTL epitopes of any arbitrary protein. To achieve this goal, Fortuna holistically integrates many different prediction tools concerning the Class I restricted antigen presentation pathway in conjunction with subsequent analysis methods. This chapter will focus on the analysis of CTL epitopes of two pathogens, HIV and Influenza A. Infection of any of these two organisms has significant impact on the human population and virulence of both have been shown to be attenuated by CTL responses. By analyzing the CTL epitope repertoire of both, insight can be attained into the general CTL response escape mutations. Furthermore, by using the clustering procedure described in Section 2.3.4 on page 48 an attempt will be made to group different sets of protein sequences together based on predicted similar CTL epitope profiles. Although only forming a part of the epitope prediction and analysis, the predictive performance result of VLTAPP will be discussed in detail and compared to other predictors. Beyond merely creating VLTAPP for prediction purposes, the author would like to enlighten the reader to the construction and evaluation of a typical sequence-based ligand predictor and subsequent physiochemical properties that can be obtained by investigating the predictor itself. The results of VLTAPP will be discussed first followed by HIV and Influenza analyses.

## 4.1 Performance of VLTAPP

This section shows the performance analysis results of VLTAPP. The results aim to show how VLTAPP performs in a discrete classification and precision of ligand IC50 values. The performance analyses are divided into two parts, namely “Single Performance Measurements” in Section 4.1.1 and “Fragmented Predictor Performance Plots” in Section 4.1.1. The former describes the performance of VLTAPP in the context of the entire data set while the latter shows how well the predictor performs with the inclusion of increasing amounts of data. This measurement is necessary when considering improving the predictor with inclusion of possible data that may be available in the future. Lastly in Section 4.1.2 it will be shown how certain physiochemical

Table 4.1: List of Performance Assessments done on VLTAPP

| Measurement  | Description  |
|--|--|
| Area of the Receiver Operatic Characteristic Curve ( $A_{ROC}$ ) | Overall performance in terms of discriminating between classes at various thresholds. Ranges from [0..1] with 1 denoting a perfect prediction.   |
| Matthew's Correlation Coefficient (MCC)                          | Another measurement of the classification performance. Ranges from [-1..1] with 1 being a perfect prediction.  |
| Residual Error   | Measurement of the quantitative performance. A value of 0 denotes a perfect prediction.  |
| Regression analysis  | A linear model is fitted between the predicted and test data. The resultant $R^2$ value is used as a performance measurement and ranges from [0..1] with 1 being a perfect prediction. |
| Correlation  | Measurement of the linear trend between training and testing data. Ranges from [0..1] with 1 being a perfect linear correlation.   |

properties of the TAP dimer binding sites can be elucidated by analyzing the weights of the trained neural network.

#### 4.1.1 Performance Measurements

Extensive evaluation was performed on the predictor. The limited data set does not allow for a perfect evaluation, nevertheless the measurements are comparable to those that exist in other studies. Both classification performance and regression performance was analysed. Performance measurements were split into two groups, namely for the entire set and concerning only the samples of longer than nine amino acids. Most other studies focus on the accuracy of a given predictor for 9-mer peptides, but as stated before, most ligands encountered by TAP are not 9-mers long. Performance measurements are listed in Table 4.1.

For single performance measurements on the overall set, a 10-fold cross validation was performed. Briefly, the training set is divided into ten subsets. With each training of the ANN, one of the ten subsets are removed and the performance of the resultant network is tested on it. This is done for each of the subsets and the results are ultimately agglomerated to make a single performance measurement. For the second group (longer than nine amino acids) a 72-fold cross validation was performed as fluctuations in performance with resampling is high for small sets. Assessment of the predictor's performance with the inclusion of more data is in addition to the single performance measurement an effective way to approximate possible future improvements in prediction by inclusion of more data. This has been illustrated in literature and is referred to as Fragmented Prediction Performance Plots (Carugo, 2007).

The procedure starts with randomly shuffling the training set and reserving a predefined

Table 4.2: VLTAPP Single performance measurement for all entries.

| Measurement                       | Value | Notes   |
|-----------------------------------|-------|---|
| Correlation (Spearman)            | 0.83  | Values for Pearson and Kendall are 0.83 and 0.65 respectively   |
| Linear Regression $R^2$           | 0.685 | The gradient for the regression line is 0.870   |
| $A_{ROC}$ Value                   | 0.942 |   |
| Matthew's Correlation Coefficient | 0.773 | The value was measured at the 2100nM threshold. A maximum MCC value of 0.784 was observed at a threshold of 2800nM. |
| Residual Error                    | 1.62  | Measured as the average difference between predicted and experimental values  |

portion of the data as testing data. The training set starts at a defined amount and increases with every iteration. Fluctuation of performance measurements is expected and observed so the procedure is repeated numerous times, randomly sampling the set with each cycle. With the resultant prediction versus test values, performance measurements are made. The plots for the results are then constructed by taking the average result per training set size. The fraction of training data ranged from 10% to 95% of the data with 50 increasing steps and resampling 30 times.

Only single performance measurements were made as FPPPs would produce results that are too convoluted. The result of performance measurements are shown and discussed in Section 4.1 on page 74. Although not shown here, it was found that the network performed optimal with approximately 300 training cycles, using six hidden neurons.

### Single Performance Measurements

The results of the single performance measurements are summarised in Tables 4.2 and 4.3 for all samples and those of ligands longer than nine amino acids respectively. The first set of performances to be analysed will be for the total set. Correlation is relatively high at 0.83 and comparable to 0.88 in a study by Bhasin et al (2003). Regression analysis revealed an R-squared value of 0.685 which is higher than the 0.614 value reported for method by Peters. The  $A_{ROC}$  value of 0.94 is high, indicating that approximately 94% of the predictions made as binder or non-binder would be correct. This value is lower than the 0.96 reported for a method by Zhang. Although no comparison to another study can be made, the Matthew's Correlation Coefficient of 0.773 is high for classification. To the author's knowledge, no residual error was measured for any other predictor in the literature. The value of 1.62 for a set with values approximately normally distributed between 3.85 and 23.61 is moderately good.

It is immediately clear that the performance of VLTAPP for predicting  $\log_2 IC_{50}$  values of longer ligands is lower. This is not surprising, given the little amount of data available for

Table 4.3: VLTAPP Single performance measurement for Ligands longer than nine amino acids.

| Measurement                       | Value | Notes  |
|-----------------------------------|-------|--|
| Correlation (Spearman)            | 0.72  | Values for Pearson and Kendall are 0.70 and 0.52 respectively  |
| Linear Regression $R^2$           | 0.481 | The gradient for the regression line is 0.722  |
| $A_{ROC}$ Value                   | 0.942 |  |
| Matthew's Correlation Coefficient | 0.498 | The value was measured at the 2100nM threshold. A maximum MCC value of 0.717 was observed at a threshold of 16500nM. |
| Residual Error                    | 2.27  | Measured as the average difference between predicted and experimental values   |

the longer ligands. However, overall classification performance seems to be high and even on par with 9-mers with an  $A_{ROC}$  value of 0.942. On the other hand, the Matthew's Correlation Coefficient of 0.659 is lower than for the overall set. A possible reason for this is the severely skewed distribution of  $\log_2 IC_{50}$  values for longer peptides. Only 29% of the longer ligands are binders. The MCC value reaches its peak at approximately the median of the set, with a value of 0.784. The linear model estimated during both regression analyses indicate that VLTAPP has a tendency to over-estimate  $IC_{50}$  values owing to the fact that both linear models have gradients of less than 1. This over-estimation is marginally higher in the prediction of longer peptides. Visual representation of the ROC curves are in Figure 4.1 on page 80 .

The only comparison to other tools that can be made is on the regression level with Peters' method. The R-squared value for regression for VLTAPP is 0.56 which is higher than the 0.48 reported for Peters' method. Having said that, it would be erroneous to immediately conclude that VLTAPP's performance is the superior, since the training sets were different. The author had more examples of longer TAP ligands. While this should inherently increase the performance of VLTAPP over Peters' method, it should be noted the distribution of lengths in the set were also different. A summary of the differences can be observed in Table 4.4 on the following page. Although the Peters' method performed well, most of the test predictions were made on ligands only one or two amino acids longer than a 9-mer. Only 16% of the Peters set was longer than 11 amino acids compared to 50% for the VLTAPP set.

### Fragmented Predictor Performance Plots of VLTAPP

Results for this method are shown in Figure 4.2 on page 81. From all the Figures, it is evident that there is a dramatic increase in performance up to approximately 50 samples. Performance then almost plateaus out, however not completely. It is not hard to deduce why the dramatic increases in performance occurs mostly when increasing the size of initially small samples; the input parameters are good enough for the ANN to quickly estimate the appropriate weights of

the nodes. Having said that, other factors may influence the performance:

1. Inappropriate input parameters
2. Insufficient dataset diversity
3. Experimental error

Since a relatively good performance is achieved with small amount of data, the author does not think the input parameters are inappropriate. It is possible the dataset is not diverse enough, given there are only 343 samples. If this is true, increasing the amount of data for ANN training from this set will have no significant improvement on the ANN's performance. Performance measurements could also be influenced by this as testing on data virtually the same as the training set is irrelevant. Still, the author concludes that the plateau is largely due to experimental error. Extrapolating IC50 for values beyond the functional parameters will have a significant effect on the accuracy of the measurements. This is also seen in MHC ligand affinity experiments where a lot of the entries have a value like  $< 1$  nM. To explain, the difference between 2000 nM and 2050 nM is marginal. The one is only 1.025 times larger than the other. On the other hand, the fold difference between 0.5 nM and 1.0 nM is 2.0. In VLTAPP terms, this means one  $\log_2$  difference.

VLTAPP's performance increase does not flatten entirely and a small upward trend can still be observed at the extreme upper ends of training sample size. The nature of the increase is too marginal on this scale to make appropriate extrapolations. Various extrapolation methods were tested and most seem to get near perfect prediction at around 1500 samples, which is obviously incorrect. The predictors performance will at best, with infinite amount of samples be at the mercy of the original experimental errors.

Inherited from experimental is the difference in measurements between two experiments. Taking MHC ligand affinity measurements as an example, Peters et al. (2006) note that the Buus and Sette datasets had a correlation of 0.65. With larger sets, this lack of perfect correlation

Table 4.4: Comparison of Length Distribution of longer TAP ligands between VLTAPP and Peters' set

| Length of Peptide | Number in Peters set | Number in VLTAPP set |
|-------------------|----------------------|----------------------|
| 10                | 36                   | 28                   |
| 11                | 18                   | 8                    |
| 12                | 6                    | 4                    |
| 13                | 1                    | 2                    |
| 14                | 0                    | 3                    |
| 15                | 1                    | 18                   |
| 16                | 1                    | 7                    |
| 17                | 0                    | 2                    |
| 18                | 1                    | 0                    |
| <b>Total</b>      | <b>64</b>            | <b>72</b>            |

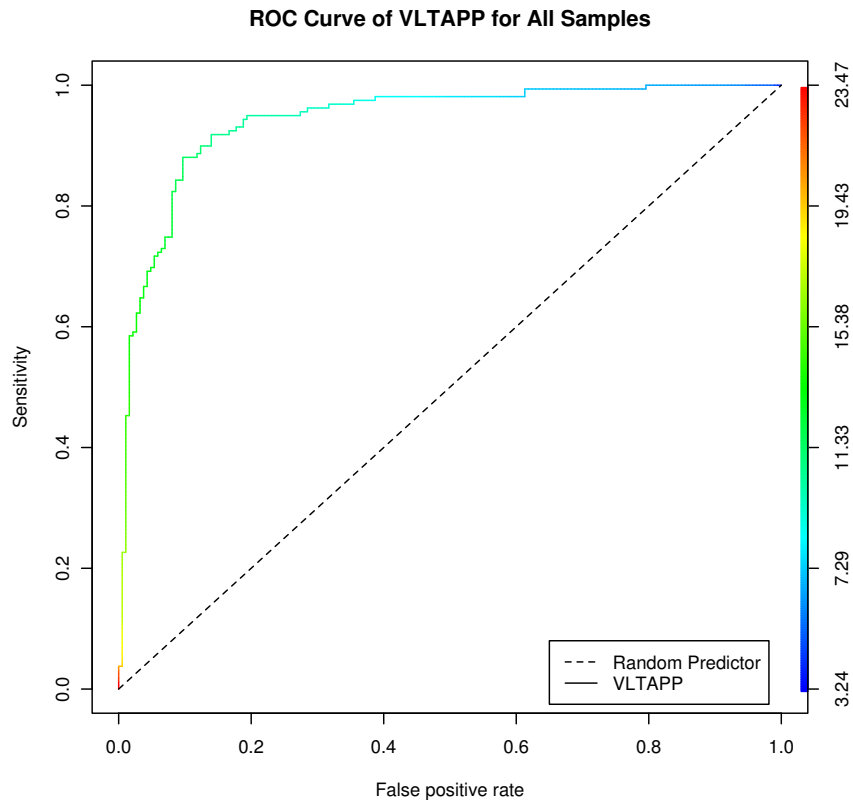


Table 4.5: Top End FPPP values vs Single Performance Measurements.

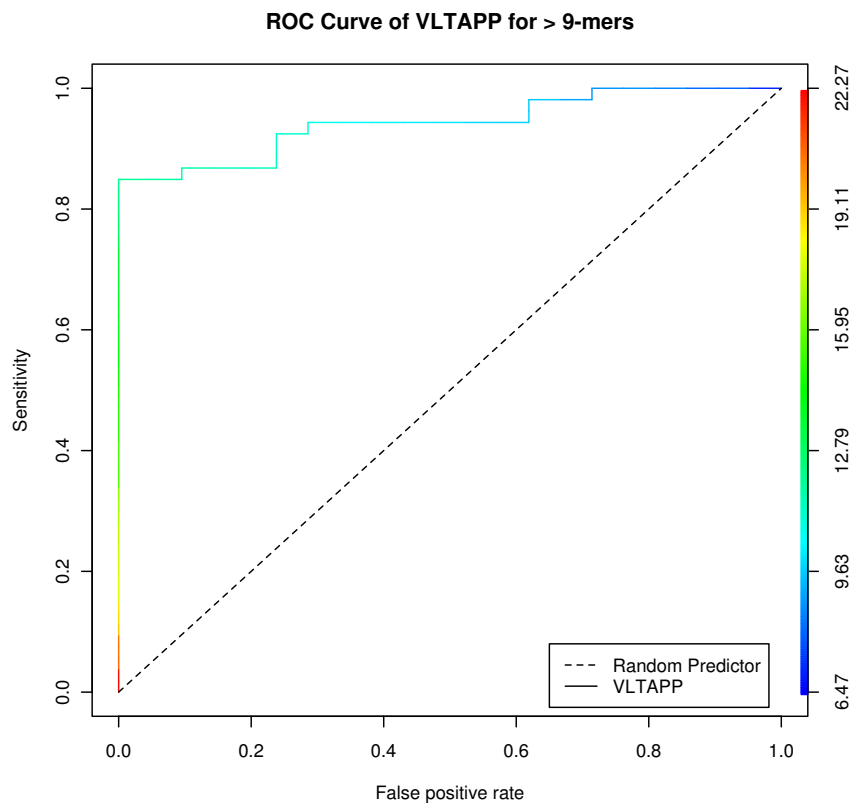
| Measurement    | Single Performance | FPPP  | Extrapolation ( $df=30$ ) |
|----------------|--------------------|-------|---------------------------|
| $A_{ROC}$      | 0.942              | 0.927 | 0.938                     |
| MCC            | 0.773              | 0.729 | 0.758                     |
| R-Squared      | 0.685              | 0.649 | 0.692                     |
| Residual Error | 1.62               | 1.74  | 1.55                      |

does not seem to influence overall prediction when combining two sets much. For 343 samples, the matter is quite different, especially considering the longer peptides for which only 72 are available.

The prudence of resampling the FPPP tests is evident in the plots of Figure 4.2. To compensate for any variance that may occur due to the small testing size of 34 for each resampling, the predicted/test values at each sample size from each resampled set were agglomerated. This procedure worked well, as can be seen from the Figures for  $A_{ROC}$ -, MCC and Linear Regression R-squared measurements (Figures 4.2a, 4.2b and 4.2d). At very small sample sizes, the prediction quality is virtually random, which is expected from training any predictor with little data. The top end values reflect the single performance measurements well and are compared in Table 4.5. The smooth-spline function in  $R$  was also used to make extrapolations to 343 samples. Interestingly, predictions for classification performances were under-estimated while performances for quantitative performances were over-estimated by this extrapolation method. This illustrates clearly that more data is needed to make FPPP from which good extrapolations can be made.



(a) ROC Curve for all samples



(b) ROC Curve for longer than 9-mer sample

Figure 4.1: Comparison of VLTAPP ROC curves for all samples and those longer than 9-mers.

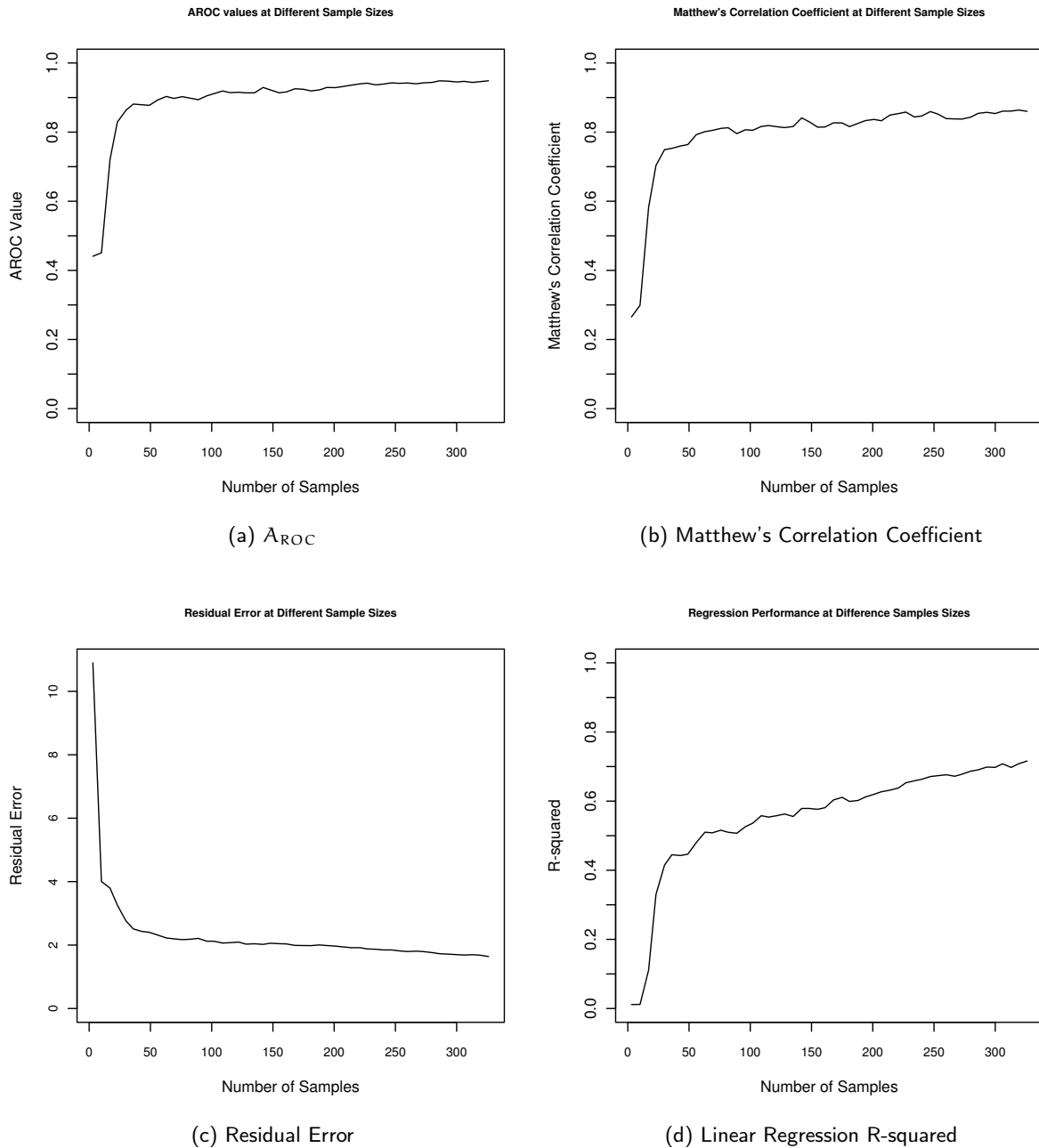


Figure 4.2: FPPP analysis using for parameters on VLTAPP. The plots were constructed from measurements of the agglomerated test data for each sample at a given sample size.

### 4.1.2 Weight Analysis of VLTAPP Artificial Neural Network

As stated before, one of the purposes of constructing a ligand binding prediction tool is to elucidate some of the relevant parameters, such as physiochemical properties that affect binding. To achieve this, maximum and minimum values for each input node was fed into VLTAPP and the resultant outputs tallied. The properties were then grouped based on position in the input and a barplot was constructed as shown in Figure 4.3 on the following page. The sign of the results were reversed, since the predictor estimates  $\log_2 IC_{50}$  values and positive weights would essentially mean a negative effect on binding. The plot depicts the total influence, or weight, of each property at a given position. The bottom graph shows the total influence that changes at a certain position has. The bottom graph shows the total  $\log_2$  influence of amino acid substitutions at a given position. The same applies for the average inputs and length. The result ties well with previous investigation on the nature of the TAP-ligand motif. The first C-terminal residue has by far the biggest influence on the score. As stated by Peters *et al.* (2003), the three N-terminal residues do contribute significantly to the score. It should be noted that the residue at C-3 also has significant weight on it's own, averaging between the N-2 and N-3 residues. The C-2, C-4 and N-4 positions have the least amount of influence on the score, but still have a few  $\log_2$  value effect on it. The average properties for the entire peptde also has far less significant impact on the output score when compared to C-1, N-1, N-2 and N-3 residues, but higher performance values were obtained with inclusion of it. Even though the total effect of the *average* and *length* inputs seem to contribute less to prediction than the other values, it is only because they consist of less input parameters than the parameters for the terminal inputs; only six parameters are used for the average values compared to nine for the terminal inputs. The top graph clearly shows that the length parameter has a profound effect on the output score.

Looking closer at the weights in the top graph, very little contradictions are observed. For instance, at the C-1 position, a high positive weight is given for basic amino acid residues while a high negative weight is given for acidic amino acids. The opposite is true for positions C-2, C-4 and N-4 where generally opposing parameters have similar weights. These inputs are therefore not as powerful as the rest of the parameters on output estimation and are also downweighed during ANN training. For the specific influences of amino acids on binding, refer to Table 4.6 on page 84.

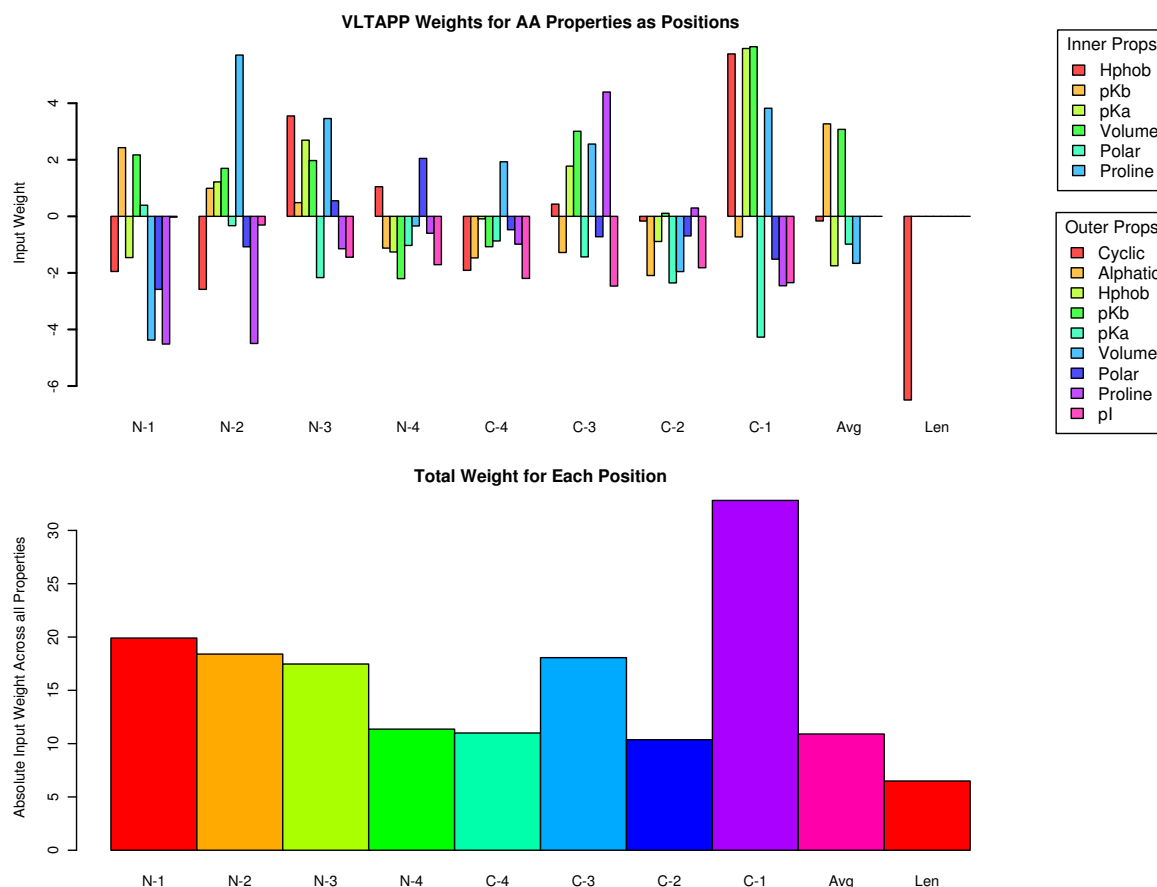


Figure 4.3: The two bar graphs above reveal the relative weights of the inputs for VLTAPP. Weights are calculate based on the influence of the maximum allowed value for the given parameter minus the minimum allowed value across all inputs. For instance, the maximum length of a peptide in the training set was 17, which translates to an input value of  $\log_2 17 = 4.087$ . The top graph reveals the relative weights of the properties at the position marked on the x-axis. The *N* and *C* letters indicate the terminal end of the peptide and the numbers correspond to the position relative to its terminal. The *Avg* value is the average value for a given property over the entire peptide. The *Len* value is the  $\log_2$  length of the peptide. Properties and their associated colours are indicated on the right. The bottom graph illustrates the total contribution of each position to the predicted score. This is calculated by the absolute sums for each property at the given position. The x-axis is labeled as in the top graph. Note that the colours do not correspond to the properties of the top graph and merely exist to provide some clarity.

Table 4.6: The table shows influence of amino acid substitutions on the predicted  $\log_2\text{IC}_{50}$  score when substituted into a reference peptide, AAAAAAAAAA. Preferred amino acids are those that produce top-end scores and disfavoured amino acids are those producing scores at the lower end.

| Amino Acid            | N-1                    | N-2        | N-3              | C-1                          |
|-----------------------|------------------------|------------|------------------|------------------------------|
| C                     | -3.26                  | -0.93      | -0.10            | 1.63                         |
| D                     | -5.11                  | -2.16      | -2.45            | -5.26                        |
| E                     | -4.53                  | -2.64      | -2.82            | -5.58                        |
| F                     | -5.69                  | -2.79      | 1.14             | 2.48                         |
| G                     | -3.59                  | -4.66      | -3.02            | -4.53                        |
| H                     | -5.81                  | -2.95      | 0.02             | 0.02                         |
| I                     | -1.17                  | 1.50       | 1.20             | 1.89                         |
| K                     | -2.77                  | -2.25      | -1.04            | 0.74                         |
| L                     | -1.12                  | 1.35       | 1.04             | 1.59                         |
| M                     | -1.76                  | 1.53       | 0.96             | 1.22                         |
| N                     | -6.16                  | -2.37      | -2.01            | -4.16                        |
| P                     | -9.03                  | -9.01      | -0.54            | -0.89                        |
| Q                     | -6.56                  | -1.93      | -1.76            | -3.85                        |
| R                     | -2.85                  | -2.03      | -0.25            | 2.58                         |
| S                     | -2.58                  | -2.49      | -1.88            | -1.58                        |
| T                     | -3.21                  | -1.94      | -1.49            | -1.08                        |
| V                     | -3.53                  | 0.37       | 0.49             | -0.16                        |
| W                     | -9.63                  | -1.83      | 0.73             | 1.85                         |
| Y                     | -5.66                  | -0.99      | 1.79             | 4.15                         |
| <b>Preferred AA</b>   | I, L, M                | V, M, L, I | H, L, M, V, W, Y | C, F, H, I, K, L, M, R, W, Y |
| <b>Disfavoured AA</b> | D, F, H, N, P, Q, W, Y | G, P       | G, N             | D, E, G, N, Q                |

## 4.2 Analysis of HIV-1 and Influenza A CTL epitopes

To assess the usefulness of *Fortuna* in predicting CTL epitopes, the epitope profiles for two distinct viral pathogens were analyzed. Both pathogens harbour experimentally determined CTL epitopes restricted to a variety of HLA allotypes. The crucial difference between the two viruses, is genetic variability of the sequences containing CTL epitopes as well as the nature of the infection itself. Influenza, being acute, does not remain in the host for too long before being eliminated or causing the death of the host, meaning that mutations in the CTL regions have little time to accumulate. HIV, being a chronic infection, has more opportunity for mutations within the regions containing CTL epitopes for the particular host to occur. As such, the priority of certain analyses on the two pathogens were different. For example, in Influenza A, clustering of sequences is useful in determining seasonal flus with highly similar CTL epitope repertoire, or conversely, grouping together sequences that are distant in terms of their CTL epitope repertoires. For HIV, being the more genetically variable of the two pathogens, SeqLogos of potential epitopes could provide insight into regions within epitopes that promote CTL escape.

### 4.2.1 Methods

#### HIV-1 Analysis

**Acquisition of Sequences and Parameter Setting** The author decided to use HIV-1 Subtype C sequences from South African patients as test subject for analysis. The highest incidence of HIV infection occurs in southern Africa and an overwhelming majority of the pathogens are of subtype C. The sequences were obtained from the LANL HIV Sequences database (<http://www.hiv.lanl.gov>). The search interface allows for changing various searching parameters. The parameters and choice by the author are shown in Table 4.7 on the following page. The important criteria to note are inclusion of sequences that have HLA information of the patient and whether the person is drug naïve. Knowing the HLA information of the patient allow us to perform analysis on only sequences of patients with a particular HLA allotype, for example, the sequences from patients with an HLA molecule of allotype *A\*0201*. Secondly, the internal structural proteins of HIV are partitioned into two groups, the *gag* group and the *pol* group. Since proteins of the *pol* group, i.e. HIV protease, Reverse Transcriptase and Integrase as well as the capsid protein of the *gag* group have drugs directed against them, it is therefore crucial to choose sequences from patients naïve to the drug treatment so that resistance mutations within these regions do not interfere with mutations relevant to CTL escape. A similar reason is given for the exclusion of the surface proteins, *gp41* and *gp120*. Both are under immunological pressure, but mainly due to escape from high affinity antibodies which most likely will interfere with the CTL epitope profile.

Table 4.7: Search Parameters for selecting HIV sequences

| Parameter   | Choice                    |
|---|---------------------------|
| Organism  | HIV-1                     |
| Subtype   | C                         |
| Include only sequences with patient HLA information | Yes                       |
| Include only full-length sequences                  | Yes                       |
| Genomic regions                                     | <i>pol</i> and <i>gag</i> |
| Include only sequences from drug naïve patients     | Yes                       |

*Fortuna* is designed to handle a very similar set of sequences and ideally, analysis of one protein sequence. To circumvent this limitation of ‘one protein only’, the sequences of the individual proteins of *gag* and *pol* were ligated with a 20-mer sequence of tryptophan (W). The predictions needed to be altered, so that the ligated regions do not yield any incorrect MHC ligands or proteasomal cleavage sites. The sequences were ligated as they appear in the *gag* and *pol* regions with the *gag* proteins preceding the *pol* proteins. A total set of 147 sequences and 189 sequences for *pol* and *gag* respectively were obtained from LANL. The *gag* and *pol* sequences with the same patient identification number and year in the FASTA title line were joined and sequences containing non-standard amino acids were filtered out, leaving a final set of 125 *gag-pol* sequences.

To determine which HLA allotypes predictions should be made for, the HLA allotypes embedded in the FASTA line were tallied and a list of the top occurring HLA allotypes were made. Concerning analysis of sequences, the sequences are from patients with HLA allotypes for which most of the supertypes are known. NetMHC only has a limited set of HLA allotypes for which ligands can be predicted. If the patient possesses an allotype that does not occur directly in the list of predictable HLA allotypes, the representative of its supertype is chosen as a substitute for MHC affinity prediction. In Table 4.8 on the next page, the frequency, supertype information and choice of HLA Allotype for which predictions of the listed allotype will be made, is shown. Where no direct match of HLA allotype occurs, the alternative HLA allotype for which predictions will be made is underlined. The choice of alternative HLA allotype was made with the aid of HLA supertype information obtained from the literature (Sidney *et al.*, 2008). The length of MHC ligands was chosen to be 9-11 amino acids. MHC threshold was set to 500 nM, proteasomal threshold set to 0.8 and maximum length of TAP ligands chosen to be 18.

**Analysis of Predictions performed on HIV Sequences** Though this study is not focused around evaluating the performances of prediction tools utilised, it is important to evaluate which



Table 4.8: HLA Allotype Frequencies and Prediction Choice

| HLA Allotype | Count | Supertype | Predicted HLA Allotype |
|--------------|-------|-----------|------------------------|
| B*5802       | 35    | B58       | <u>B*5801</u>          |
| A*6802       | 26    | A2        | A*6802                 |
| A*3001       | 25    | A1/A3     | A*3001                 |
| B*1510       | 25    | B39       | <u>B*3901</u>          |
| B*4201       | 24    | B7        | <u>B*3501</u>          |
| B*1503       | 24    | B27       | <u>B*2705</u>          |
| A*02         | 23    | A2        | A*0201                 |
| A*2301       | 16    | A24       | A*2301                 |
| B*0801       | 14    | B8        | B*0801                 |
| A*0301       | 13    | A3        | A*0301                 |
| A*3002       | 11    | A1        | A*3001                 |
| B*5801       | 11    | B58       | B*5801                 |
| A*74         | 11    | A3        | <u>A*0301</u>          |
| B*44         | 11    | B44       | <u>B*4401</u>          |

experimentally determined CTL epitopes can be predicted by using Fortuna. The Los Alamos National Library provides a compendia of CTL epitopes defined for many HLA allotypes (Korber *et al.*, 2007). The compendium also provides a list of optimally defined epitopes. One of the criteria for an ‘optimal epitope’ is that it exhibits a dominant response, meaning that escape mutations in this epitope has a significantly negative effect on immune response. Optimal epitopes for the HLA allotypes in Table 4.8 associated with supertypes A2, A3, B7 and B58 were noted and searched for in the pathway prediction results performed on the previously mentioned HIV protein sequences. Because different protein sequences were artificially joined, the positions of the experimentally determined epitopes needed to be remapped. The sequences were searched for sequences of optimally defined epitopes and the positions noted. In certain cases, the exact sequence for a determined epitope did not occur in any of the testing HIV sequences. In this event, the sequences were searched for the closest matching sequence word by comparing regions within the sequences with the epitope sequence based on a BLOSUM35 matrix. If no reasonable match could be found, the epitope was omitted from the study. All the prediction results were tested for being positive at the positions where the epitopes occur. Though information such as proteasomal cleavage and TAP affinity are not always noted for the listed epitopes, it is assumed that these prediction results need at least be above a reasonable threshold to be considered as a positive result.

**Epitope Entropy and Frequency Analysis** The epitopes predicted for HIV sequences will be analyzed for both frequency and sequence entropy. The attempt here is to demonstrate whether the mutations within a particular epitope are geared towards providing CTL escape by lowering the affinity of the epitope to a TCR or whether the mutations abrogate the ability

of the epitope sequence to bind to MHC. In the event of the former, mutations should occur in non-anchor positions and possibly in positions that are relevant to MHC-epitope and TCR interactions as noted by Frankild *et al.* (2008). For the latter, the mutations should occur more in the terminal ends of the epitope and frequency of the epitope and its variants above the MHC binding threshold should be lower. The SeqLogo's generated by the Epitope/Frequency analysis should also provide insight into which amino acids are 'favoured' by CTL escape mechanisms. The author notes that the favouring of a certain amino acid does not necessarily only reflect on CTL escape, but also maintaining viral fitness. Only the epitope sequence profiles of the optimally defined epitopes will be considered.

**Cluster Analysis of HIV Sequence Immunological Profiles** It may seem senseless clustering sequences of HIV from unrelated patients together, but clustering of sequences from patients with shared HLA allotypes could provide insight into patterns of escape mutations. Groups of sequences with similar distances to other sequences likely share a similar immunological state that could not necessarily be determined by direct phylogenetic analysis performed on aligned sequences alone. By determining which groups are immunologically close/distant to each other from the heatmap produced in cluster analysis, the in-depth comparison (by using the compare sequences tool) could provide insights into the epitope sequences that are vulnerable to the effect of mutations therein. The sets of optimally defined epitopes for each HLA allotype will be analysed separately and only one sequence containing the relevant HLA allotype. Using sequences from patients that do not possess the required HLA allotype could lead to false conclusions, because either the epitope sequence remains conserved due to lack of immunological challenging or mutations exist due to overlapping epitopes that are expressed by the patient's HLA allotypes.

Parameters for the clustering procedure were set to the HLA allotype to be tested, lengths of the ligands in question, position where the ligands occur, MHC threshold of 1000 nM. The effects of using prediction results as weights was also tested by first clustering with the prediction results as weights and then without. In all cases, the distance of a missing epitope was set to 0.3. It should be noted that HLA allotypes that belong to the same supertype were tested together for the purpose of increasing the sample size of the clustering procedure.

**Self-Epitope Discovery** To determine if there are potential self-epitopes within the HIV sequences, all the predicted epitopes for the HLA allotypes listed were analysed and examined for BLAST hits to human protein sequences. The reason for not limiting the analysis to OTEs was to test whether there are indeed potential epitopes that are not immunologically active due to their similarity to ligands presented by cells in the human body. As stated by Frankild, it is very likely to miss self-epitopes by using a BLOSUM matrix based comparison method, so

it is expected that only a fraction of the epitopes predicted would be similar to self-epitopes, if any. Epitopes with high self-similarity were tested for conservation as well as the frequency of sequences containing the potential self-epitope of variants of it with sufficient MHC binding affinity. The list of predicted self-epitopes will be compared with epitopes of the same or similar sequence obtained in the literature based on immunogenicity.

### Influenza A Analysis

**Acquisition of Sequences** The Los Alamos National Laboratory also provides another site dedicated to Flu research (<http://www.flu.lanl.gov>). The site contains a comprehensive list of influenza sequences, both protein and nucleotide based. Since H1N1 and H3N2 are the more common serotypes of Influenza A to infect humans, sequences from these two serotypes were obtained. The search interface allows limiting the search to specific proteins. The search was limited to all the proteins of Influenza A except the envelope proteins, Hemagglutinin and Neuraminidase, based on similar reasons used for limiting HIV sequences. HA and NA are susceptible to humoral immunity and mutations within these protein sequences to evade antibody responses may occur in regions containing CTL epitopes and would give a wrong impression on the nature of CTL escape mutations. The source of the sequences were set to anywhere in the world and from any year. The resultant set of sequences are a mix of different protein sequences from different genes. The sequences were grouped together based on the origin of the sequence as per standard Influenza nomenclature, i.e. Virus Type/Geographic Origin/Strain Number/Year of Infection (Virus Subtype). The sequence groups that do not contain all of the protein sequences PA, PB1, PB2, NS1, NS2, M1, M2 and NP were excluded. The sequences were joined on the same principle as the HIV sequences, including a 20-mer tryptophan sequence between the proteins. The final set contains 4402 sequences. *Fortuna* is not meant to handle more than approximately 200 sequences at a time, so the final sequence set was trimmed by the following procedure:

1. Isolate sequences for a specific Influenza A serotype
2. Extract sequences from different years with at most 3 sequences from a single year

This gave a representable set of Influenza sequences for each serotype over years, some from as far back as the 1930s. This procedure was not applied to the 2009 pandemic strain of H1N1, where 45 sequences were extracted. The H1N1 sequences from the pandemic were treated separately from the other H1N1 sequences, meaning that the list sequences from non-2009 pandemic H1N1 isolates was generated using the same aforementioned criteria. The final list contains 236 sequences and the breakdown is shown in Table 4.9.

Table 4.9: Breakdown of Influenza A Sequence Set

| Serotype                 | Count      |
|--------------------------|------------|
| H1N1 (non-2009 pandemic) | 74         |
| H1N1 (2009 pandemic)     | 41         |
| H3N2                     | 80         |
| H2N2                     | 41         |
| <b>Total</b>             | <b>236</b> |

**Epitope Prediction Analysis** The analysis of epitopes will be limited to only the HLA A0201 allotype for which a comprehensive set of epitopes exist in the literature, i.e. MHC affinity and immunogenicity have been described. A portion of the predicted epitopes were tested against a list of known epitopes and the experimental values matched with prediction results where available.

**Entropy and Frequency Analysis** Entropy and frequency analysis will be performed as for HIV sequences. The difference here is that predicted epitopes that do not exist in the literature will be scrutinise for amino acid substitutions. It is assumed that internal influenza proteins, being mutationally inert when compared to HIV proteins, only acquire mutations when sufficient pressure is applied. That is, in this case, from the CTL response. Similar to the HIV CTL epitope entropy/frequency analysis, the analysis on influenza epitopes could provide insight into the conservation of epitope sequence and whether mutations are partial towards direct CTL escape by amino acid substitutions that attenuate MHC-peptide-TCR interactions or by abrogating the ability of the epitope to bind to MHC.

**Cluster Analysis** To investigate how the different Influenza group together immunologically a few factors need to be taken into account. Influenza A serotypes, being defined by Neuramidinase and Haemaggluttinin, may still share similar CTL epitopes borne from the internal proteins. Clustering will be performed on the basis of epitopes for HLA A0201 and B5801, all of which being representatives of HLA supertypes A2, B8 respectively. The motivation is that sufficient literature is available for definition of epitopes restricted to these HLA allotypes. Clustering will be performed in the context of epitopes obtained from the literature and compared to clustering performed by using all predicted epitopes.

The disproportionate nature of 2009 Pandemic Influenza sequences forced the author to exclude these sequences from the initial study, since the UPGMA clustering procedure is likely to cluster large groups of sequences together and insight could be lost in terms of relationship of 2009 pandemic flu with strains from other years. As a sub-study, the 2009 pandemic flu sequences will be tested amongst themselves to determine patterns of immune escape. The 2009 pandemic flu sequences provided a unique opportunity to test numerous flu sequences that were collected

in a comparatively short time. The clustering procedure was first performed on a naïve basis, meaning all the predicted epitopes were considered for clustering, and then based on epitopes found in the literature. This is to test the influence of possibly incorrectly predicted epitopes.

It should be noted that an appropriate criteria needs to be used to define the clusters identified. For Flu the dates when the sequences were obtained can be used, for instance, to see any relationship with modern Flu strains with past strains. For HIV, the question is a little more complex. The only criterium the author considered to aid in the definition of the clusters would be the level of HIV in the blood of the patients. Although this criterium is sound in theory, it should be noted that the level of HIV in the blood of patients does not necessarily directly correlate with the CTL epitope repertoire.

## 4.2.2 Epitope Analysis Results

### Pathway Prediction Results and Epitope Variants

When using prediction tools to analyse sequences for potential CTL epitopes, it is often surprising how many hits are found. However, the very nature of the prediction tools employed will result in many true positives and false negatives. It is already mentioned that epitopes defined in the literature will be examined in context of their variants. Having said that, it is also important to understand why the pathway prediction results cannot be used as is; meaning it is unwise to assume predicted epitopes to be true if they have not been experimentally validated. The pathway prediction results for both HIV and Influenza A were analysed in fairly the same manner. The prediction scores for the literature defined epitopes were assimilated and are displayed in various tables. The relation of the literature defined epitopes to the predicted epitopes were also determined to illustrate how well the combination of current tools involved in the Class I restricted antigen presentation pathway can determine immunodominant epitopes.

Central to the analysis of epitopes is determining where sequence variations occur. Sequence variations close to the terminal ends generally have a profound effect on the binding affinity of the epitope to MHC, whereas mutations more to the center have profound effects on interaction of the pMHC complex to the TCR of the original epitope. Only the optimal epitope sequence variants were considered. Considering the variants of all the predicted epitopes would be too exhaustive. SeqLogos were used to represent both the predicted binder and non-binder variants of an epitope, juxtaposed and placed in a table. Epitope variants of HIV are shown in Section 4.2.3.

### HIV Pathway Prediction Results

It has already been mentioned that protein products of gp160 (i.e. gp41 and gp120) are excluded from analysis, because T-Cell epitope mutations may be affected by B-Cell escape mutations. Beyond this, certain sequences that were originally planned to be included in the CTL epitope analysis were excluded as well. Some of the sequences obtained from LANL contain wildcard characters at certain positions and Fortuna does not accept sequences that contain non-standard amino acid characters. Due to the fact that all the sequences from a patient are grammatically ligated together, a whole set of potentially useful sequences may be excluded due to wildcard characters existing in some of the included protein sequences. This is especially true for Vpu that reduced the amount of legitimate sequences by a significant amount. Vpu is also a short sequence and does not contain many optimal CTL epitopes. The exception to this exclusion rule is sequences ending in a wildcard character, which was merely stripped from the sequence. Some epitopes were not found as exact matches in the sequences due to HIV's high mutability. When no match to an epitope was found, the closest matching sequence was used as a representative. In all subsequent analyses, the representative sequences were marked as being a non-match to the original sequence.

**Optimal HIV Epitope Pathway Predictions** The prediction results including proteasomal cleavage, TAP affinity, MHC affinity and immunogenicity were tested for the optimal epitopes restricted to the HLA allotypes A\*0201, A\*0301 and B\*5801. Prediction results for the aforementioned epitopes are summarised in Table 4.11 on page 95. The thresholds used for the different prediction tools were 0.1 for proteasomal cleavage, 2100nM for TAP prediction and 1000nM for MHC prediction. It should be noted that the allotype used for predictions of HLA A\*0201 restricted epitopes were made by using HLA\*A212, because NetMHC failed to predict appreciable IC<sub>50</sub> values for HLA A\*0201. The NetMHC predicted values are summarised in Table 4.10 on page 94 and it is clear that that all predictions for HLA\*A0201 epitopes are under 1000nM when using HLA\*A0201. Epitopes for HLA\*A6802 were also tested for binding affinity under HLA\*A02xx allotypes since they do belong to the same A2 supertype, but almost none of the epitopes were predicted to be “cross binders”. This factor is important when considering epitope variants across the sequences from patients with HLA allotypes that fall in the same supertype. For the A\*0301 and B\*5801 restricted epitopes almost all the predictions fell within appreciable value ranges. The prediction results shown in Table 4.11 on page 95 are shown in a format that is both relative and absolute. TAP and Proteasomal cleavage scores are the log<sub>10</sub> scores relative to the predefined threshold, i.e. the log<sub>10</sub> value *below* the IC<sub>50</sub> threshold for TAP affinity and the log<sub>10</sub> value *above* the threshold for proteasomal cleavage. The MHC score is directly defined as the IC<sub>50</sub> value in nM, while the score for POPI is the log<sub>10</sub> value of the

predicted spot forming units of the epitope. To be noted are the values in a colour other than green. Red indicates scores that did not meet the required threshold. With regards to MHC  $IC_{50}$  values, orange values are those above 500nM which is generally regarded as the cutoff point and 1000nM, the leanient threshold decided upon by the author. It is expected that optimally defined epitopes will have scores above the prediction thresholds and it can be assumed that red scores are scores incorrectly predicted. This is especially apparent for POPI scores, where the majority of the epitopes are predicted to have no or very little immunogenicity. The scores for proteasomal cleavage and TAP affinity are mostly above the threshold levels, but it should be mentioned that the proteasomal cleavage was set to a low level.

### Optimal HIV Epitopes in Context of Other Predicted Epitopes

The data in Table 4.11 on page 95 is restricted to the optimally defined HIV epitopes restricted to HLA allotypes A\*0201 (A\*212) A\*0301 and B\*5801. However, these are not the only epitopes predicted. To assess the position of the optimally defined epitopes relative to other predicted epitopes, a list of predicted epitopes using the same parameters as for the optimally defined epitopes was obtained with the added condition that at least 30% of the sequences contained an predicted epitope at a particular position. Both 9-mer and 10-mer epitopes were searched for. The resultant list contains 230 predicted epitopes across all the previously mentioned HIV proteins. The results are summarised in Table 4.14 on page 99. Only the top 30 predicted epitopes (top half) along with the optimally defined epitopes are shown (bottom half). As is evident, only one of the optimally defined epitopes were in the top 30, namely ILKEPVHGV which is at a position of 25. The prediction scores are ranked according to the combination of the Immunoproteasome score, Immunoproteasome produced TAP ligand affinity, MHC affinity and immunogenicity, i.e.  $MIPT_i$ . HIV is a chronic infection and it is assumed that cells that are infected and immunogenic would be under the influence of  $IFN-\gamma$ . Some of the predicted “Top Tier” epitopes were indeed defined in the complete list of known HIV CTL epitopes, but in most cases the research consensus (obtained from the summary available at the HIV LANL site) was that the epitopes only elicited responses in a fraction of the patients. Other top epitopes were not found in the literature and are assumed to either not exist as CTL epitopes or elicit an insignificant immune response.



Table 4.10: Predictions for many HLA Allotypes of superclass A2. The green, orange and maroon numbers indicate predicted IC50 values at or below 500nM, between 500nM and 1000nM and above 1000nM respectively. Peptides marked with <sup>+</sup> are those for which only a close match was found in the sequences and those marked with \* are peptides for which no close match was found.

| Peptide                 | Restricted Allotype | A0201 | A0202 | A0212 | A6801 | A6802 | POPI |
|-------------------------|---------------------|-------|-------|-------|-------|-------|------|
| AIIRILQQL               | A*0201              | 897   | 30    | 74    | 33627 | 1950  | 0    |
| ALVEICTEM*              | A*0201              | 29    | 12    | 98    | 34132 | 10625 | 0    |
| FLGKIWPSYK <sup>+</sup> | A*0201              | 7226  | 1185  | 412   | 404   | 32949 | 0    |
| ILKEPVHGV               | A*0201              | 372   | 77    | 3     | 38214 | 3425  | 2    |
| LVGPTPVNI               | A*0201              | 4762  | 1407  | 661   | 36667 | 2172  | 3    |
| PLTFGWCYKL              | A*0201              | 10985 | 2726  | 96    | 36869 | 20446 | 0    |
| SLYNTVATL               | A*0201              | 163   | 109   | 4     | 20944 | 11158 | 0    |
| VIYQYMDDL               | A*0201              | 695   | 287   | 18    | 24407 | 15903 | 0    |
| VLEWRFSRL               | A*0201              | 1456  | 2053  | 30    | 37165 | 17517 | 1    |
| DTWAGVEAIR              | A*6801              | 46073 | 35532 | 30557 | 32    | 23527 | 3    |
| ITKGLGISYGR*            | A*6801              | 46892 | 26388 | 21530 | 21    | 23294 | 0    |
| DTVLEEWNL <sup>+</sup>  | A*6802              | 40752 | 20571 | 15937 | 14266 | 822   | 0    |
| ETYGDTWTGV              | A*6802              | 4507  | 1081  | 22    | 1324  | 5     | 0    |
| GAETFYVDGA              | A*6802              | 25224 | 5127  | 21252 | 38363 | 26467 | 0    |
| ITLWQRPLV               | A*6802              | 799   | 10136 | 344   | 13409 | 2011  | 0    |
| IVTRIVELL <sup>+</sup>  | A*6802              | 1666  | 111   | 4762  | 8535  | 120   | 0    |



Table 4.11: Below are the results of predictions performed on epitopes obtained from the literature and the fragments from which they originate. The epitopes are grouped according to the HLA allotype to which they are restricted. The  $P_c$  and  $P_i$  abbreviations indicate the threshold adjusted score for proteasomal prediction on the C-terminal end for the constitutive and immunoproteasome respectively. The  $T_c$  and  $T_i$  values are the averaged scores for TAP ligands that originate from predicted proteasomal fragments.  $MHC_{IC50}$  is the IC50 value of the ligand and POPI the predicted immunogenicity. All prediction scores are represented as  $\log_{10}$  scores subtracted by the threshold score for that particular prediction and limited to 0.00. The the exception being  $MHC_{IC50}$ , which is given as raw IC50 values measured in nM. Epitope sequences marked with ‘+’ are sequences that closely resemble the original sequence in the literature, but for which no direct match was found in the sequences.

| Allotype | Sequence                 | $P_c$ | $T_c$ | $P_i$ | $T_i$ | $MHC_{IC50}$ | POPI | MIPT <sub>c</sub> | MIPT <sub>i</sub> |
|----------|--------------------------|-------|-------|-------|-------|--------------|------|-------------------|-------------------|
| A0212    | AIIRILQQL                | 1.16  | 0.66  | 1.31  | 0.71  | 74           | 0    | 4.48              | 4.81              |
| A0212    | FLGKIWPSQK <sup>+</sup>  | 0.80  | 0.99  | 0.90  | 1.06  | 886          | 0    | 4.14              | 4.47              |
| A0212    | LVGPTPVNI                | 0.99  | 0.92  | 1.05  | 1.00  | 661          | 3    | 7.23              | 7.56              |
| A0212    | VIYQYDDL                 | 1.01  | 0.00  | 1.41  | 0.00  | 18           | 0    | 1.74              | 2.07              |
| A0212    | ILKEPVHGV                | 1.03  | 0.27  | 1.08  | 0.36  | 3            | 2    | 6.42              | 6.75              |
| A0212    | SLYNTVATL                | 1.17  | 0.24  | 1.70  | 0.18  | 4            | 0    | 4.39              | 4.72              |
| A0212    | PLTFGWCYKL               | 0.90  | 0.00  | 1.36  | 0.00  | 96           | 0    | 1.93              | 2.26              |
| A0212    | ALTEICTEM                | 0.79  | 0.00  | 0.77  | 0.00  | 140          | 2    | 2.85              | 3.18              |
| A0212    | VLKWRFDSSL*              | 1.16  | 0.82  | 1.47  | 0.83  | 14           | 1    | 6.76              | 7.09              |
| A0301    | KTKPPLPSVSK              | 0.68  | 0.74  | 0.90  | 0.77  | 30           | 0    | 4.66              | 4.99              |
| A0301    | AVDLSFFLK                | 0.42  | 0.39  | 0.67  | 0.42  | 89           | 0    | 2.78              | 3.11              |
| A0301    | AIFQSSMTK                | 0.67  | 0.39  | 0.60  | 0.51  | 11           | 3    | 6.93              | 7.26              |
| A0301    | QVPLRPMTYK               | 0.80  | 0.03  | 0.99  | 0.08  | 158          | 1    | 2.71              | 3.04              |
| A0301    | KIRLRPGGKKK              | 1.02  | 0.86  | 1.35  | 0.87  | 83           | 1    | 5.98              | 6.31              |
| A0301    | HMYISKKAK                | 0.68  | 0.77  | 0.96  | 0.79  | 37           | 0    | 4.70              | 5.03              |
| A0301    | RLRPGGKKKY <sup>+</sup>  | 1.35  | 0.12  | 1.86  | 0.07  | 721          | 0    | 1.90              | 2.23              |
| A0301    | KIRLRPGGK                | 0.59  | 0.43  | 0.65  | 0.51  | 65           | 2    | 5.21              | 5.54              |
| A0301    | KLVDRELNK                | 0.72  | 0.67  | 0.89  | 0.72  | 111          | 0    | 3.91              | 4.24              |
| A0301    | QIYPGIKVK                | 0.96  | 0.38  | 1.16  | 0.42  | 120          | 2    | 5.15              | 5.48              |
| A0301    | GIPHPAGLK                | 0.57  | 0.32  | 0.78  | 0.35  | 55           | 2    | 4.88              | 5.22              |
| A0301    | AVFIHNFKR                | 0.78  | 0.78  | 0.98  | 0.82  | 127          | 2    | 6.26              | 6.59              |
| A0301    | ALTEICTEMEK <sup>+</sup> | 0.51  | 0.37  | 0.59  | 0.45  | 391          | 2    | 4.14              | 4.47              |
| A0301    | KMRSHTNDVK <sup>+</sup>  | 0.66  | 0.41  | 0.79  | 0.47  | 136          | 2    | 4.89              | 5.22              |
| A0301    | AVDLSFFLK <sup>+</sup>   | 0.42  | 0.50  | 0.67  | 0.53  | 89           | 0    | 3.14              | 3.47              |
| A0301    | RVKQWPLTEEK <sup>+</sup> | 0.53  | 0.00  | 0.74  | 0.00  | 254          | 0    | 0.60              | 0.93              |
| A0301    | KTKPPLPSVSK <sup>+</sup> | 0.68  | 0.47  | 0.90  | 0.50  | 30           | 0    | 3.77              | 4.10              |
| B5801    | TSTLQEQIGW               | 1.59  | 0.41  | 2.11  | 0.35  | 74           | 2    | 6.08              | 6.41              |
| B5801    | IAMESIVIW                | 1.58  | 0.00  | 1.97  | 0.00  | 10           | 0    | 2.00              | 2.33              |
| B5801    | RSLYNTVATLY              | 1.17  | 0.20  | 1.63  | 0.16  | 216          | 1    | 3.51              | 3.84              |

Table 4.12: Below are the top 30 ranked predicted HIV epitopes along with the optimally defined epitopes.

| Rank | Sequence    | Count | $P_c$ | $T_c$ | $P_i$ | $T_i$ | $MHC_{IC50}$ | POPI | $MIPT_c$ | $MIPT_i$ |
|------|-------------|-------|-------|-------|-------|-------|--------------|------|----------|----------|
| 1    | KLVPVDPREV  | 55    | 0.77  | 0.95  | 0.97  | 0.98  | 9            | 2    | 7.96     | 8.29     |
| 2    | EMMTACQGV   | 103   | 0.91  | 0.50  | 0.94  | 0.59  | 7            | 3    | 7.71     | 8.04     |
| 3    | VASGYIEAEV  | 73    | 0.70  | 0.71  | 0.75  | 0.80  | 36           | 3    | 7.51     | 7.84     |
| 4    | ELAENREIL   | 104   | 1.00  | 0.49  | 1.50  | 0.44  | 22           | 3    | 7.29     | 7.62     |
| 5    | HLKTAVQMAV  | 101   | 0.78  | 0.67  | 0.97  | 0.71  | 6            | 2    | 7.22     | 7.55     |
| 6    | KLAGRWPVKV  | 100   | 0.75  | 0.53  | 0.87  | 0.59  | 3            | 2    | 7.04     | 7.37     |
| 7    | IVTDSQYAL   | 100   | 1.11  | 0.48  | 1.60  | 0.43  | 71           | 3    | 6.84     | 7.18     |
| 8    | SLVKHHMYI   | 56    | 0.83  | 0.53  | 1.12  | 0.55  | 6            | 2    | 6.82     | 7.15     |
| 9    | VLDVGDAYF   | 106   | 0.88  | 0.60  | 1.39  | 0.54  | 169          | 3    | 6.64     | 6.97     |
| 10   | GIWQLDCTHL  | 100   | 1.09  | 0.76  | 1.26  | 0.81  | 117          | 2    | 6.57     | 6.90     |
| 11   | QLPEKDSWTV  | 63    | 0.95  | 0.27  | 1.12  | 0.32  | 3            | 2    | 6.45     | 6.78     |
| 12   | VLDVGDAYFS  | 105   | 0.71  | 0.65  | 1.01  | 0.66  | 330          | 3    | 6.35     | 6.68     |
| 13   | KLLWKGEGA   | 103   | 0.89  | 0.59  | 0.94  | 0.68  | 43           | 2    | 6.22     | 6.55     |
| 14   | RLRRYSTQV   | 44    | 0.91  | 0.85  | 0.90  | 0.95  | 4            | 0    | 6.18     | 6.51     |
| 15   | GLQRGWEAL   | 47    | 1.15  | 0.43  | 1.63  | 0.39  | 4            | 1    | 5.98     | 6.31     |
| 16   | VIQDNSDIKV  | 91    | 0.70  | 0.68  | 0.94  | 0.71  | 10           | 1    | 5.97     | 6.30     |
| 17   | QLGIPHPAGL  | 105   | 1.14  | 0.46  | 1.39  | 0.48  | 50           | 2    | 5.95     | 6.29     |
| 18   | LLWKGEGAVV  | 104   | 0.90  | 0.43  | 0.87  | 0.54  | 2            | 1    | 5.94     | 6.27     |
| 19   | ALNPGLEET   | 90    | 0.86  | 0.96  | 0.98  | 1.03  | 14           | 0    | 5.92     | 6.25     |
| 20   | KLVSSGIRKV  | 70    | 0.81  | 0.50  | 0.57  | 0.68  | 39           | 2    | 5.90     | 6.23     |
| 21   | KVGSLLQYLAL | 101   | 1.09  | 0.75  | 1.59  | 0.70  | 493          | 2    | 5.88     | 6.22     |
| 22   | RMRIRTWNSL  | 55    | 0.98  | 0.86  | 1.48  | 0.80  | 89           | 1    | 5.88     | 6.21     |
| 23   | ILKEPVHGV   | 86    | 1.03  | 0.10  | 1.08  | 0.19  | 3            | 2    | 5.86     | 6.19     |
| 24   | YMDDLTVGS   | 102   | 0.87  | 0.51  | 0.96  | 0.58  | 5            | 1    | 5.86     | 6.19     |
| 26   | SLCLFSYHRL  | 36    | 1.23  | 0.33  | 1.54  | 0.34  | 37           | 2    | 5.76     | 6.09     |
| 27   | RTQDFWEVQL  | 101   | 1.19  | 0.53  | 1.59  | 0.51  | 196          | 2    | 5.67     | 6.00     |
| 28   | LVSSGIRKV   | 70    | 0.81  | 0.50  | 0.57  | 0.68  | 69           | 2    | 5.66     | 5.99     |
| 29   | QIYPGIKVRQ  | 38    | 0.6   | 0.72  | 0.62  | 0.81  | 261          | 2    | 5.59     | 5.92     |
| 30   | AEWDRLHPV   | 73    | 0.62  | 0.53  | 0.82  | 0.57  | 62           | 2    | 5.58     | 5.91     |
| 25   | ILKEPVHGV   | 86    | 1.03  | 0.10  | 1.08  | 0.19  | 3            | 2    | 5.86     | 6.19     |
| 39   | LVGPTPVNI   | 93    | 0.99  | 0.35  | 1.05  | 0.43  | 661          | 3    | 5.34     | 5.67     |
| 53   | VLKWRFDSSL  | 1     | 1.16  | 0.28  | 1.47  | 0.29  | 14           | 1    | 4.97     | 5.30     |
| 55   | ALTEICTEM   | 2     | 0.79  | 0.38  | 0.77  | 0.49  | 140          | 2    | 4.91     | 5.24     |
| 62   | SLYNTVATL   | 58    | 1.17  | 0.38  | 1.70  | 0.32  | 4            | 0    | 4.86     | 5.19     |
| 112  | VIYQYMDDL   | 105   | 1.01  | 0.36  | 1.41  | 0.34  | 18           | 0    | 3.96     | 4.29     |
| 141  | PLTFGWICYKL | 28    | 0.90  | 0.49  | 1.36  | 0.45  | 96           | 0    | 3.55     | 3.88     |
| 159  | AIIRILQQL   | 26    | 1.16  | 0.31  | 1.31  | 0.37  | 74           | 0    | 3.33     | 3.66     |
| 223  | FLGRIWPSHK  | 7     | 0.82  | 0.27  | 0.90  | 0.34  | 772          | 0    | 1.82     | 2.15     |

### Influenza Pathway Prediction Results

All of the preselected protein sequences for Influenza were included in the pathway analysis and none had to be removed.

**Optimal Influenza Epitopes** The pathway prediction results for Influenza epitopes restricted to the HLA\*A0201, HLA\*A1101, HLA\*B0702 and HLA\*B4002 are listed in Table 4.13 on the following page. Again, it can be observed that the least accurate results are found in immunogenicity prediction. Overall, the results for the other predictions are reasonable, i.e. the values represent optimal epitopes. The epitopes are displayed in the same way as the HIV epitopes in Table 4.11 on page 95. All the optimal epitopes could be predicted by their respective HLA allotypes to which they are restricted. Therefore, an HLA prediction “substitute” as was needed for certain HLA\*A0201 restricted epitopes of HIV is not necessary for any of the Influenza epitopes.

**Optimal Epitopes in the Context of other Predicted Epitopes** The top 30 predicted HLA\*A0201 restricted epitopes are listed in Table 4.13 on the following page. Five out of eighteen of the optimal epitopes from the literature are included in the top 30 results and twelve fall in the upper half of the 180 predicted epitopes. The results are significantly better for Influenza than HIV epitopes. Interestingly, according to the Immune Epitope Database and Analysis Resource (<http://www.immuneepitope.org>), the 2nd best predicted epitope ILGFVFTLTV is elsewhere defined in the literature as immunogenic, although not listed in the source the author considered. However, the fourth top predicted epitope, GMITQFESL was negative in eliciting an immune response, although it was shown to bind to HLA\*A0201.

Table 4.13: Below are predicted values for FLU CTL epitopes from the literature. The results are expressed in the same way as in Table 4.11.

| Allotype | Sequence   | $P_c$ | $T_c$ | $P_i$ | $T_i$ | MHCIC50 | POPI |
|----------|------------|-------|-------|-------|-------|---------|------|
| A0201    | GILGFVFTL  | 0.34  | 0.00  | 0.87  | 0.00  | 18      | 3    |
| A0201    | CLPACVYGL  | 0.50  | 0.99  | 1.02  | 0.91  | 61      | 2    |
| A0201    | FQGRGVFEL  | 0.53  | 0.98  | 1.07  | 0.89  | 78      | 3    |
| A0201    | NMLSTVLGV  | 0.21  | 0.77  | 0.37  | 0.85  | 10      | 0    |
| A0201    | FMYSDFHFI  | 0.20  | 0.85  | 0.53  | 0.85  | 2       | 0    |
| A0201    | FQVDCFLWHV | 0.21  | 0.77  | 0.32  | 0.87  | 9       | 0    |
|          |            |       |       |       |       |         |      |
| A1101    | ASCMGLIYNR | 0.19  | 0.00  | 0.67  | 0.00  | 17      | 2    |
| A1101    | RLFFKCIYRR | 0.05  | 1.19  | 0.21  | 1.26  | 93      | 0    |
| A1101    | SVQPAFSVQR | 0.30  | 1.08  | 0.62  | 1.09  | 30      | 3    |
| A1101    | SVQRNLPFER | 0.25  | 0.83  | 0.63  | 0.80  | 18      | 3    |
| A1101    | KLVGINMSKK | 0.00  | 0.00  | 0.04  | 0.13  | 63      | 2    |
| A1101    | GTFEFTSFFY | 0.62  | 0.67  | 0.85  | 0.71  | 9       | 0    |
| A1101    | SFSFGGFTFK | 0.00  | 0.81  | 0.23  | 0.86  | 32      | 2    |
| A1101    | VLRGFLILGK | 0.00  | 0.00  | 0.01  | 0.14  | 214     | 0    |
| A1101    | KFLPDLYDYK | 0.00  | 0.00  | 0.20  | 0.06  | 393     | 0    |
|          |            |       |       |       |       |         |      |
| B0702    | LPFDRTTVM  | 0.48  | 0.00  | 0.49  | 0.00  | 8       | 3    |
| B0702    | SPIVPSFDM  | 0.28  | 1.06  | 0.27  | 1.16  | 230     | 0    |
| B0702    | QPEWFRNVL  | 0.52  | 0.75  | 1.02  | 0.70  | 27      | 2    |
| B0702    | QPEWFRNIL  | 0.41  | 0.78  | 0.99  | 0.70  | 58      | 2    |
|          |            |       |       |       |       |         |      |
| B4002    | TEVETYVLSI | 0.15  | 0.00  | 0.27  | 0.02  | 293     | 1    |
| B4002    | SEQAAEAMEV | 0.22  | 1.08  | 0.51  | 1.09  | 612     | 0    |
| B4002    | GERQNANEI  | 0.12  | 0.87  | 0.46  | 0.87  | 602     | 0    |
| B4002    | QEIRTFQFL  | 0.36  | 0.85  | 0.42  | 0.93  | 479     | 2    |

Table 4.14: Below are the top 30 ranked predicted Flu epitopes along with the optimally defined epitopes.

| Rank | Sequence    | Count | $P_c$ | $T_c$ | $P_i$ | $T_i$ | $MHC_{IC50}$ | POPI | $MIPT_c$ | $MIPT_i$ |
|------|-------------|-------|-------|-------|-------|-------|--------------|------|----------|----------|
| 2    | ILGFVFTLTV  | 235   | 0.82  | 0.72  | 0.91  | 0.79  | 56           | 3    | 7.17     | 7.5      |
| 4    | GMITQFESL   | 185   | 1.09  | 0.65  | 1.6   | 0.6   | 70           | 3    | 7.11     | 7.44     |
| 5    | YMFESKSMKL  | 117   | 1.03  | 0.67  | 1.51  | 0.63  | 9            | 2    | 7        | 7.33     |
| 6    | MQFSSLTVNV  | 230   | 0.92  | 0.52  | 0.94  | 0.62  | 28           | 3    | 6.91     | 7.24     |
| 7    | KIYKTYFEKV  | 87    | 0.83  | 0.78  | 1.08  | 0.8   | 20           | 2    | 6.81     | 7.14     |
| 8    | GQMSRPMFL   | 75    | 1.16  | 0.63  | 1.68  | 0.57  | 150          | 3    | 6.78     | 7.11     |
| 9    | FINEQGESIV  | 91    | 0.72  | 0.75  | 0.75  | 0.84  | 145          | 3    | 6.76     | 7.09     |
| 10   | FVANFSMEL   | 236   | 0.89  | 0.33  | 1.47  | 0.26  | 20           | 3    | 6.39     | 6.72     |
| 12   | LLQNSQVYSL  | 129   | 0.98  | 0.61  | 1.28  | 0.62  | 24           | 2    | 6.32     | 6.65     |
| 13   | RVMVSP LAV  | 199   | 0.88  | 0.59  | 0.78  | 0.72  | 203          | 3    | 6.24     | 6.57     |
| 15   | RQMVAATTNPL | 71    | 0.79  | 0.58  | 1.04  | 0.6   | 212          | 3    | 6.1      | 6.43     |
| 16   | FSMELPSFGV  | 235   | 0.9   | 0.32  | 0.97  | 0.4   | 37           | 3    | 6.09     | 6.42     |
| 17   | MMWEINGPES  | 227   | 0.88  | 0.81  | 0.88  | 0.91  | 187          | 2    | 5.99     | 6.32     |
| 19   | GLKDDLENL   | 143   | 1.07  | 0.4   | 1.58  | 0.35  | 156          | 3    | 5.91     | 6.24     |
| 20   | ILTSESQLTI  | 188   | 0.84  | 0.45  | 0.94  | 0.52  | 143          | 3    | 5.88     | 6.22     |
| 21   | KLSDYEGRL   | 139   | 1.08  | 0.63  | 1.29  | 0.67  | 113          | 2    | 5.84     | 6.17     |
| 22   | RLNKRSYLI   | 73    | 0.77  | 0.74  | 1.19  | 0.71  | 143          | 2    | 5.77     | 6.1      |
| 23   | QMSRPMFLYV  | 75    | 0.99  | 0.45  | 1.27  | 0.47  | 33           | 2    | 5.68     | 6.01     |
| 24   | VIFDRLETL   | 127   | 1.14  | 0.47  | 1.5   | 0.46  | 54           | 2    | 5.65     | 5.98     |
| 25   | FVSHKEIESV  | 185   | 0.91  | 0.81  | 1.26  | 0.8   | 49           | 1    | 5.61     | 5.95     |
| 26   | WLIEEVRHRL  | 124   | 1.06  | 0.29  | 1.3   | 0.32  | 145          | 3    | 5.57     | 5.9      |
| 27   | SLPGHTNEDV  | 62    | 0.74  | 0.47  | 0.99  | 0.49  | 270          | 3    | 5.57     | 5.9      |
| 28   | AQDVIMEVV   | 236   | 0.87  | 0.74  | 1.14  | 0.75  | 301          | 2    | 5.55     | 5.88     |
| 29   | GISSMVEAMV  | 225   | 0.97  | 0.37  | 1.06  | 0.44  | 230          | 3    | 5.53     | 5.86     |
| 30   | ILVRGNPSPV  | 229   | 0.7   | 0.46  | 0.74  | 0.54  | 300          | 3    | 5.45     | 5.78     |
| 1    | GMFNMMLSTV  | 236   | 0.82  | 0.57  | 0.73  | 0.7   | 11           | 3    | 7.39     | 7.72     |
| 3    | RMQFSSSLTV  | 230   | 0.89  | 0.67  | 0.99  | 0.74  | 48           | 3    | 7.15     | 7.48     |
| 11   | GILGFVFTL   | 235   | 1.04  | 0.26  | 1.57  | 0.2   | 18           | 3    | 6.37     | 6.7      |
| 14   | CLPACVYGL   | 46    | 1.2   | 0.63  | 1.72  | 0.57  | 61           | 2    | 6.21     | 6.54     |
| 18   | FQGRGVFEL   | 156   | 1.23  | 0.28  | 1.77  | 0.22  | 78           | 3    | 5.97     | 6.3      |
| 38   | FMYSDFHFI   | 236   | 0.9   | 0.57  | 1.23  | 0.57  | 2            | 0    | 5.12     | 5.45     |
| 51   | FQVDCFLWHV  | 146   | 0.91  | 0.6   | 1.02  | 0.67  | 9            | 0    | 4.65     | 4.98     |
| 59   | LLMDALKLSI  | 209   | 0.81  | 0.53  | 1.11  | 0.54  | 7            | 0    | 4.46     | 4.79     |
| 62   | RLIDFLKDV   | 236   | 0.85  | 0.48  | 1.18  | 0.48  | 56           | 1    | 4.41     | 4.74     |
| 74   | NMLSTVLGV   | 236   | 0.91  | 0.49  | 1.07  | 0.54  | 10           | 0    | 4.24     | 4.57     |
| 77   | ALLKHF E I  | 236   | 0.71  | 0.51  | 0.72  | 0.6   | 80           | 1    | 4.19     | 4.52     |
| 99   | SMIEAESSV   | 212   | 0.79  | 0.45  | 0.77  | 0.55  | 18           | 0    | 3.74     | 4.07     |
| 115  | NLYNIRNLHI  | 234   | 0.97  | 0.06  | 1.21  | 0.08  | 239          | 2    | 3.48     | 3.81     |
| 121  | MLLRSAIGQV  | 72    | 0.96  | 0.45  | 1.04  | 0.53  | 65           | 0    | 3.36     | 3.69     |
| 127  | FLEESHPI    | 231   | 0.88  | 0.34  | 1.05  | 0.38  | 22           | 0    | 3.35     | 3.68     |
| 150  | RLIDFLKDVM  | 197   | 0.81  | 0.35  | 0.81  | 0.45  | 462          | 1    | 3.02     | 3.35     |
| 165  | WMMAMKYPI   | 194   | 0.6   | 0.19  | 0.92  | 0.2   | 16           | 0    | 2.74     | 3.07     |
| 180  | CLLQSLQIQI  | 236   | 0.83  | 0.34  | 0.92  | 0.41  | 307          | 0    | 2.18     | 2.51     |

### 4.2.3 Variants of Epitopes

#### Variants of HIV CTL Epitopes

Although many sequences are available to test for sequence variants of the epitopes, only the sequences from patients to which the optimal defined epitopes are restricted to were considered for determining sequence variants. The epitopes and respective variants for HLA\*A0201 are listed in Table 4.15 and HLA\*B5801 in Table 4.16. The Tables show the epitope sequence from the literature, the frequency of occurrence within patients with the appropriate HLA allotype, the averaged entropy over the length of the sequence, the sequence variants and finally sequence logos representing the sequences that are predicted to be presented on MHC and those that aren't. It is immediately apparent that mutations within most of the epitope sequences for both HLA A\*0201 and B\*05801 are confined to specific positions. This is in accordance with the literature where the impact of mutations within the SLYNTVATL epitope (hereafter referred to as SL9) on immunogenicity was assessed. Indeed, all the variants listed in Table 4.15 correspond to those reported by the researchers. It should be noted that SL9 also exists within an epitope presented by HLA B\*5801, namely RSLYNTVATLY (hereafter referred to as RL11). Interestingly the mutations of SL9 correspond with the mutations of RL11. This is not too surprising, as the more central part of the epitope is recognised by the appropriate TCR. Still, the similarity in the proportion of SL9:Y3->F3 and RL11:Y4:F4 as well as SL9:T8->V8 and RL11:T9->V9 is significant, considering the sequence logos were created from sets of sequences that only overlap by 30%. The sequence logos also reveal to a certain extent whether the mutations have an effect on the epitopes MHC binding ability or interaction between the peptide-MHC complex and the TCR. For example, FLGRIWPSHK has the majority of its mutations occurring at position 9, which is the position of the anchor residues, having mostly an effect on peptide-MHC interaction, whereas VLKWEFDSSL, the mutations are more focused on positions 3 and 5. This is indicative of mutations that cause interference with MHC-peptide-TCR interaction.

It is plausible that novel epitopes could be obtained from the list of predicted epitopes other than the optimally defined epitopes on the basis of sequence variability. However, it is a common occurrence for epitopes from different HLA allotypes to overlap and mutations within a predicted epitope may only exist due to selective pressures applied to a flanking epitope. This also applies to the reverse argument, where a predicted epitope may be excluded outright based on the lack of mutational variability, but one of the HLA\*A0201 restricted ODEs investigated, VIYQYMDDL, showed in the context of the sequences examined, very little variability. Surprisingly, some predicted epitopes with very little sequence variability, had poor combination scores, which does raise the confidence in the prediction results somewhat.

**Association of Epitope Variants with Levels of HIV** To investigate whether there is a direct relationship between epitope variants and blood levels of HIV, the variants of the p17 related epitope, SLYNTVATL. This epitope has been studied in detail before and it was established that this epitope is immunodominant and subject to mutation and that these mutations do have an effect on levels of HIV (Iversen *et al.*, 2006). The sample size of HIV sequences from patients with the HLA allotype A\*02 (*sic*) is 28. However, as shown before, the HLA\*B5801 restricted epitope RSLYNTVATLY shows similar patterns of variation. Thus, using the sequences obtained from patients with the B\*58 allotype brings up the sample size to 52. The results are shown in 4.4 on page 103. Although no direct correlation between the presence of SLYNTVATL and HIV levels, it can be deduced that the variants of the epitope are associated with the more extreme HIV viral loads.

Table 4.15: Variants for certain HLA\*0201 restricted optimal epitopes. The epitope sequence from the literature is shown in the first column. The second and third column show the frequency of MHC binders (i.e. all the epitope variants that have sufficient binding affinity to the HLA molecule) and the the averaged entropy of the sequences. Nominal sequence and its variants are shown in column four. Column five and six show the sequence variants in SeqLogo format.

| Lit Sequence | Frequency | Entropy | Sequences   | Binder SeqLogos | Non-binder SeqLogos |
|--------------|-----------|---------|---|-----------------|---------------------|
| ALTAICEEM    | 0.66      | 0.06    | ALTAICEEM (10)<br>...E.... (5)<br>...E...D... (1)<br>...E...K... (1)<br>...I... (1)<br>...G.... (1)   |                 |                     |
| VIYQYMDL     | 1.00      | 0.00    | VIYQYMDL (29)   |                 | None                |
| ILKEPVHGV    | 1.00      | 0.02    | ILKEPVHGV (24)<br>.....A (2)<br>..R..... (2)<br>..Q..... (1)  |                 | None                |
| SLYNTVATL    | 1.00      | 0.03    | SLYNTVATL (18)<br>..F..... (7)<br>.....V. (2)<br>..F...V. (1)<br>.....I... (1)  |                 | None                |
| LVGPTPVNI    | 0.93      | 0.01    | LVGPTPVNI (25)<br>.I..... (2)   |                 |                     |
| AIIRILQQL    | 1.00      | 0.11    | AIIRILQQL (7)<br>.LL..... (5)<br>.L..T.... (2)<br>.LM..... (1)<br>.S.V..... (1)<br>..K..... (1)<br>.L..L.... (1)<br>.L..M.... (1)<br>TLL..... (1)<br>.M..T.... (1)<br>.M.V..... (1)<br>VLL..T... (1)<br>TLT..... (1)<br>TL..M.... (1)<br>M.T..... (1)<br>.L.M..... (1)<br>..T.T.... (1) |                 | None                |
| PLTFGWCFKL   | 0.97      | 0.05    | PLTFGWCFKL (20)<br>.....Y.. (5)<br>..Y..... (1)<br>.....P... (1)<br>.....R... (1)   |                 |                     |
| VLKWEFDSSL   | 0.97      | 0.12    | VLKWEFDSSL (6)<br>.M.K.... (4)<br>...V.... (3)<br>...Q.... (3)<br>...K.... (3)<br>...K...Q. (1)<br>..Q.R...L. (1)<br>..Q.Q.... (1)<br>..R.....N. (1)<br>..R..... (1)<br>..R.K.... (1)<br>..M.K...G. (1)<br>..Q.K...I.. (1)<br>..Q.KL... (1)   |                 |                     |
| FLGRIWPSHK   | 0.38      | 0.07    | FLGRIWPSHK (3)<br>...K...Q. (2)<br>...K...R. (2)<br>...F.... (1)<br>...K...S. (1)<br>...KV.... (1)<br>...KL.... (1)   |                 |                     |



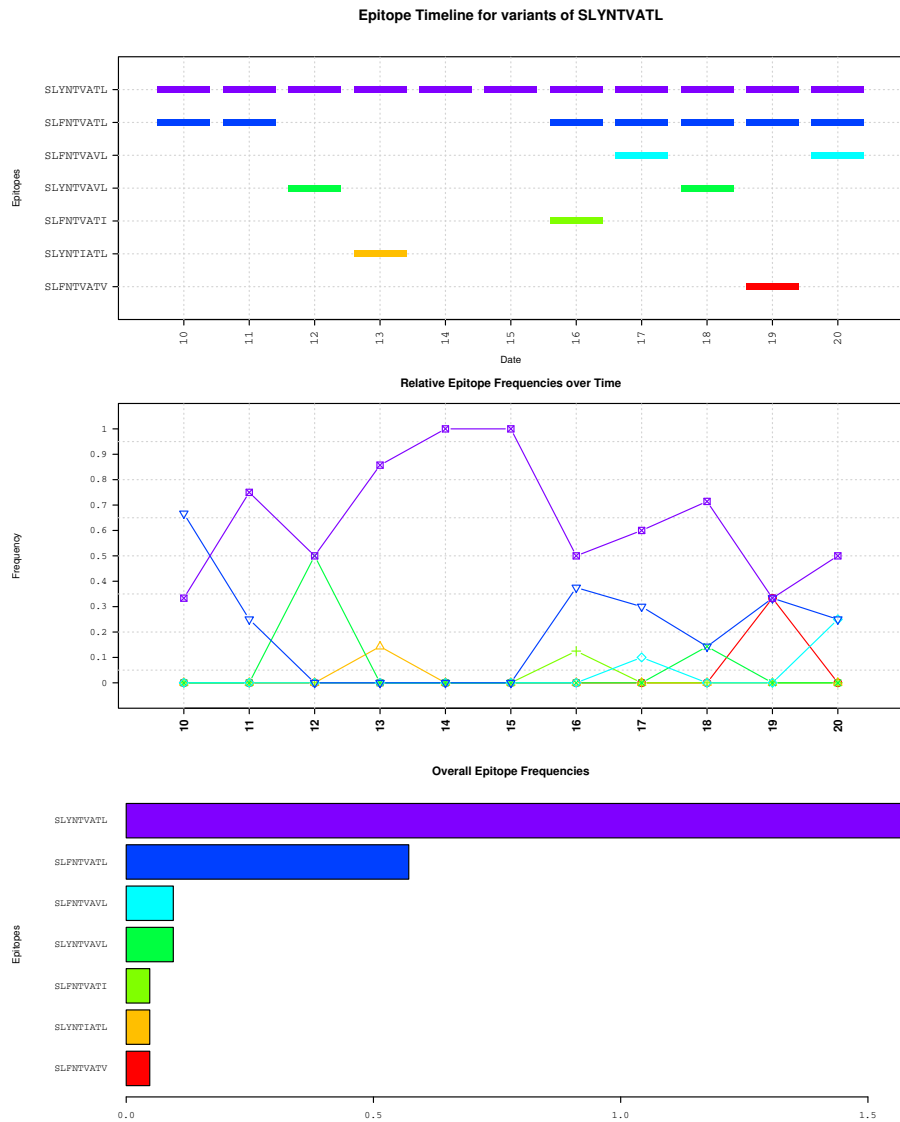


Figure 4.4: The figure depicts the variants of SLYNTVATL. The top graph maps the epitope variants to the  $\log_2\text{RNA}/\text{mL}$  values of the patients the sequences were obtained from. The middle graph depicts the relative frequencies of epitope variants found at the indicated viral level. The bottom graph shows the relative frequencies of all the variants of SLYNTVATL.

Table 4.16: Entropy/Frequency

| Lit Sequence | Frequency | Entropy | Sequences  | SeqLogo (B) | SeqLogo(NB) |
|--------------|-----------|---------|--|-------------|-------------|
| IALESIVIV    | 1.00      | 0.06    | IALESIVIV (36)<br>..M..... (30)<br>..I..... (5)<br>..T..... (5)<br>..V..... (3)<br>..S..... (3)<br>..Q..... (3)<br>..SM..... (3)<br>..G..... (2)<br>V..... (2)<br>..T..... (2)<br>..Q.C..... (2)<br>V.I..... (1)<br>M.M..... (1)<br>..K..... (1)<br>..M..I... (1)<br>..M.AV... (1)<br>V.M..... (1)<br>.....V. (1)<br>..W..... (1)<br>..S..... (1)<br>..R..... (1)<br>..TT..... (1)<br>VVM..... (1) |             | None        |
| KSLYNTVATLY  | 0.97      | 0.04    | KSLYNTVATLY (32)<br>R..... (24)<br>R..F..... (17)<br>...F..... (14)<br>.....V... (4)<br>..F...V... (4)<br>.....I.V... (1)<br>R..H..... (1)<br>R..F...V... (1)<br>R.....F (1)<br>R..F...V. (1)<br>..F.L..... (1)<br>.....I.V.W (1)<br>...F.....I. (1)<br>R.....V... (1)<br>R..F..I... (1)   |             |             |
| TSTLQEQIAW   | 0.99      | 0.03    | TSTLQEQIAW (67)<br>.....T. (11)<br>..N..... (6)<br>.....V... (6)<br>.....G. (4)<br>.....Q. (3)<br>...A..... (2)<br>..S..... (2)<br>..S.....G. (1)<br>..N.....N. (1)<br>..N...V... (1)<br>..N.G..... (1)<br>....D..... (1)<br>..N.....T. (1)  |             |             |

### Variants of Influenza CTL Epitopes

The variants of optimal Influenza epitopes were described separately for serotypes H1N1 and H3N2. Only variants for HLA\*A0201 restricted epitopes were considered. Variants for H1N1 and H3N2 epitopes are listed in Tables 4.17 on the following page and 4.18 on page 107 respectively. Only epitopes with at least one variant found in the sequences tested are included in the Tables. The reason for the separate analysis of epitope variants between H1N1 and H3N2 is that the author noticed discrepancies between the epitopes of the two serotypes. Most notably is the absence of the epitopes **FQGRGVFEL** in the H3N2 sequences analyzed. Overall the sequence conservation of CTL epitopes for both serotypes is high with top-end epitopes having an entropy of 0.02 – 0.03. Some of the optimal epitopes were only found in a fraction of the sequences, meaning they were predicted with MHC IC<sub>50</sub> values of > 500 nM. That is, they were predicted as non-binders. The author also noted that some of the predicted epitopes have high entropy associated with them. This begs the question whether these are in fact epitopes with high enough immunogenicity that makes mutations within these regions beneficial in terms of escaping immune surveillance. Although putative inferences can be made from the predicted results, experimental validation is necessary to positively or negatively identify these regions as epitopes. Nevertheless, some of these epitopes for H3N2 are shown in Table 4.19 on page 107.

Table 4.17: Variants for certain HLA\*0201 restricted optimal epitopes of Influenza A H1N1.









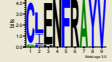
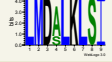

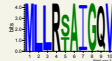
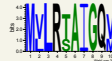
| Lit Sequence | Frequency | Entropy | Sequences   | Binder SeqLogos   | Non-binder SeqLogos   |
|--------------|-----------|---------|---|---|---|
| FQGRGVFEL    | 0.92      | 0.01    | FQGRGVFEL (68)  |    |    |
| FQVDCFLWHV   | 1.00      | 0.01    | FQVDCFLWHV (69)<br>.....I (5)   |    | None  |
| RLNKRGYLI    | 0.98      | 0.05    | RLNKRGYLI (42)<br>.....S... (37)<br>K.D...S... (25)<br>K...S... (5)<br>K.T...S... (1)<br>K..R.... (1)<br>K..RK... (1)<br>...SS... (1) |    |    |
| WMMAMKYPI    | 1.00      | 0.02    | WMMAMKYPI (73)<br>.....R... (41)<br>...V.... (1)  |  | None  |
| RMQFSSLTV    | 1.00      | 0.01    | RMQFSSLTV (110)<br>.....F.. (4)<br>G.....I (1)  |  | None  |
| SLENFRAYV    | 0.44      | 0.04    | SLENFRAYV (48)<br>.....T.. (1)<br>.K..... (1)<br>N..... (1)   |  |  |
| LMDALKLSI    | 1.00      | 0.01    | LMDALKLSI (88)<br>..S..... (27)   |  | None  |
| LLMDALKLSI   | 1.00      | 0.01    | LLMDALKLSI (88)<br>..S..... (27)  |  | None  |
| MLLRSATGV    | 0.95      | 0.03    | MLLRSATGV (57)<br>...T..... (42)<br>...I..... (3)<br>.....L.. (2)<br>.....H. (2)<br>...N.... (1)<br>..I..... (1)<br>.....P. (1)       |  |  |

Table 4.18: Variants for certain HLA\*0201 restricted optimal epitopes of Influenza A H3N2.

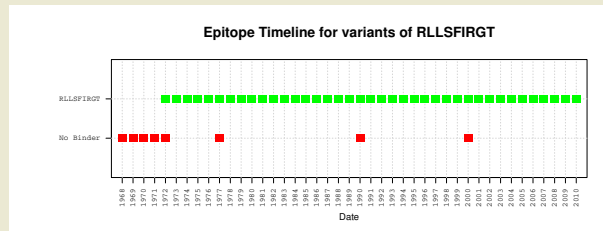
| Lit Sequence | Frequency | Entropy | Sequences   | Binder SeqLogos | Non-binder SeqLogos |
|--------------|-----------|---------|---|-----------------|---------------------|
| KLDKRSYLI    | 0.03      | 0.02    | KLDKRSYLI (1)<br>RMN..... (1)                               |                 |                     |
| CLESFRAYV    | 0.01      | 0.00    | CLESFRAYV (1)   |                 |                     |
| LLRSAISQV    | 0.01      | 0.02    | LLRSAISQV (1)   |                 |                     |
| MLLRSAIGQI   | 0.40      | 0.02    | MLLRSAIGQI (30)<br>.....S.V (1)<br>.....L.V (1)             |                 |                     |
| CLLQSLQQI    | 1.00      | 0.00    | CLLQSLQQI (80)  |                 | None                |
| SMIEAESSV    | 1.00      | 0.01    | SMIEAESSV (64)<br>.....I (14)<br>.V..... (1)<br>.I..... (1) |                 | None                |
| GILGFVFTL    | 1.00      | 0.00    | GILGFVFTL (80)  |                 | None                |
| FQVDCFLWHI   | 1.00      | 0.02    | FQVDCFLWHI (46)<br>.....V (33)<br>.....C.V (1)              |                 | None                |
| NMLSTVLGV    | 1.00      | 0.00    | NMLSTVLGV (80)  |                 | None                |
| FMYSDFHFI    | 1.00      | 0.00    | FMYSDFHFI (80)  |                 | None                |

Table 4.19: Variants for certain HLA\*0201 restricted predicted epitopes of Influenza A H3N2 with high sequence entropy.

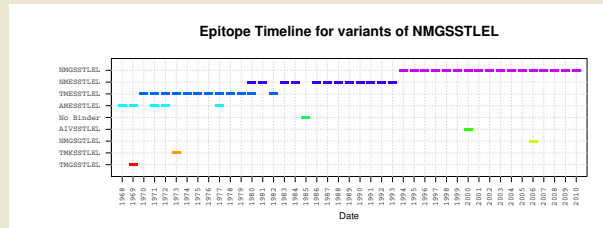
| Lit Sequence | Frequency | Entropy | Sequences   | Binder SeqLogos | Non-binder SeqLogos |
|--------------|-----------|---------|---|-----------------|---------------------|
| KLDKRSYLI    | 0.03      | 0.02    | KLDKRSYLI (1)<br>RMN..... (1)                               |                 |                     |
| CLESFRAYV    | 0.01      | 0.00    | CLESFRAYV (1)   |                 |                     |
| LLRSAISQV    | 0.01      | 0.02    | LLRSAISQV (1)   |                 |                     |
| MLLRSAIGQI   | 0.40      | 0.02    | MLLRSAIGQI (30)<br>.....S.V (1)<br>.....L.V (1)             |                 |                     |
| CLLQSLQQI    | 1.00      | 0.00    | CLLQSLQQI (80)  |                 | None                |
| SMIEAESSV    | 1.00      | 0.01    | SMIEAESSV (64)<br>.....I (14)<br>.V..... (1)<br>.I..... (1) |                 | None                |
| GILGFVFTL    | 1.00      | 0.00    | GILGFVFTL (80)  |                 | None                |
| FQVDCFLWHI   | 1.00      | 0.02    | FQVDCFLWHI (46)<br>.....V (33)<br>.....C.V (1)              |                 | None                |
| NMLSTVLGV    | 1.00      | 0.00    | NMLSTVLGV (80)  |                 | None                |
| FMYSDFHFI    | 1.00      | 0.00    | FMYSDFHFI (80)  |                 | None                |

**Timeline Analysis of Predicted Epitopes** To investigate whether the predicted epitopes with high frequency are associated with the immunogenic evolution of Influenza, timelines were constructed. The timelines are a visualization of epitopes at a specific positions occurring at the times the sequences were obtained. That is, it shows the variants of epitopes at different time points. Of special interest were the sequences obtained in and around 1968 when the Hong Kong flu pandemic struck. This pandemic was associated with the H3N2 serotype. H3N2 is also the current seasonal flu. In Figure 4.5 on the next page the timeline profiles of selected HLA\*A0201 restricted epitopes are shown. Although none contain any of the optimal epitopes, these epitopes did present some interesting results. For all of them, a particular variant or set of variants occurred during the Hong Kong flu pandemic of ca. 1968. The optimal epitopes, being generally highly conserved, did not show any difference at the time points associated with the pandemic. It could be pure coincidence that these variants occurred at said timepoints, however the author notes that having at least six epitopes with variants (or in the case of SIWIELDEI, absence) reduces the probability of coincidence. The PB1 epitope, SMDKEEIEI has also recently reverted to a variant seen from 1968–1977. Predicted epitopes scores are shown in Figure 4.6 on page 110. The figures represent the sum of the predicted scores for the epitopes, meaning that the presence or absence of an epitope influences the score. For the optimal epitopes, there is a clear upward trend in the sum of the scores, however a marginal downward trend is observed for the average scores. Taking all the predicted scores into account, there is a definitive downward trend for the sum of the epitope scores and a similar downward trend in the average predicted scores. Again, this poses the question whether Influenza is steadily losing HLA\*A0201 immunogenic CTL epitopes. However, the author urges the reader not to find too much comfort in these results as they are, afterall, predicted. As can be seen in Figure 4.6e on page 110, there seems to be no relation between the time sequences were obtained and total HLA\*A1101 restricted epitope scores for predicted.

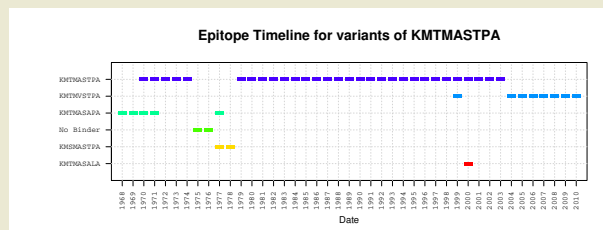
Figure 4.5: Variants of selected epitopes over time. The labels of the Figures indicated the Influenza in which the epitope exists.



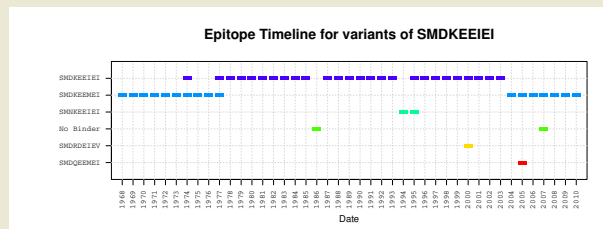
(a) NP



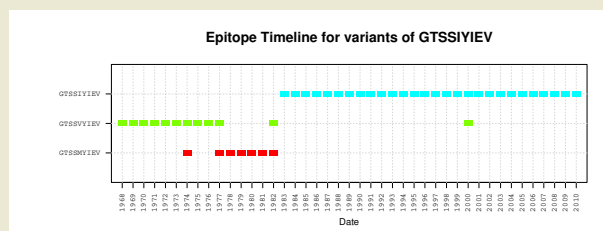
(b) NP



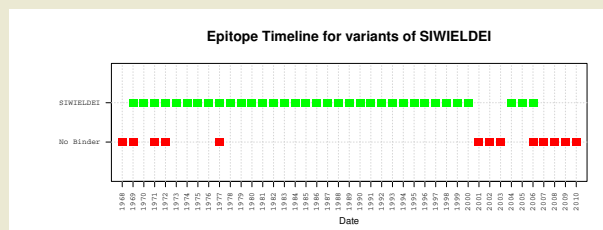
(c) NS1



(d) PB1



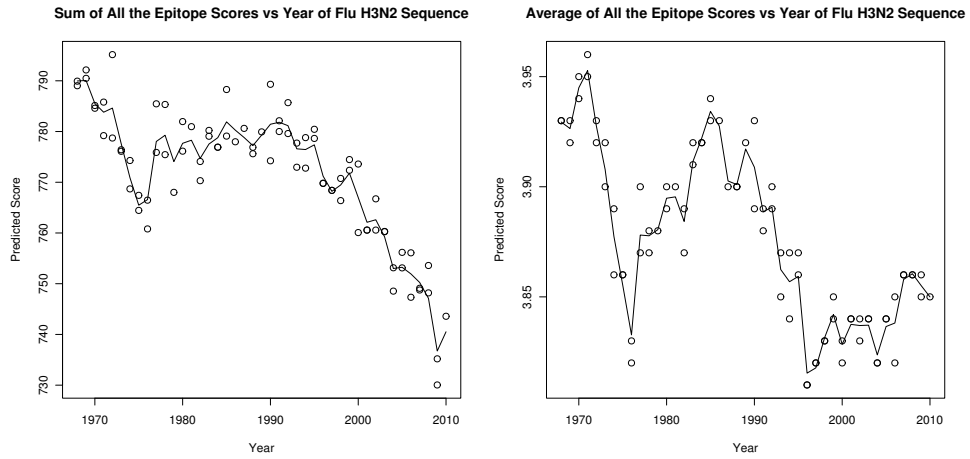
(e) PB2



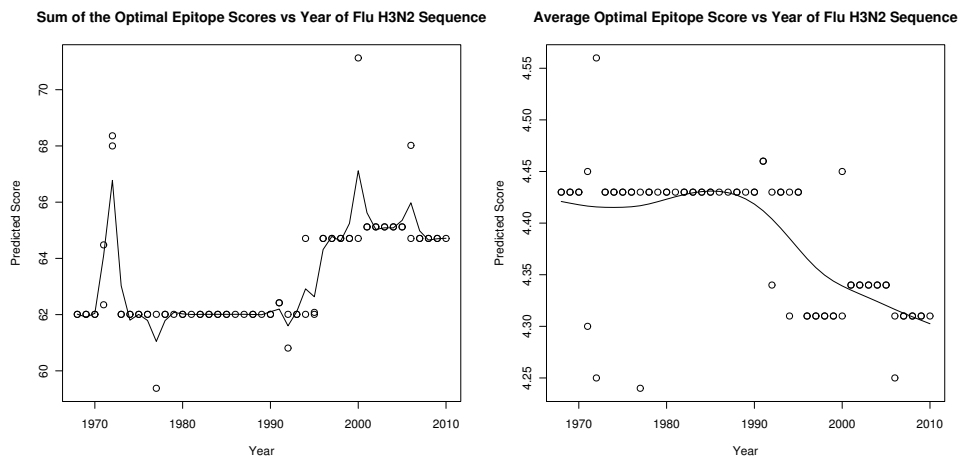
(f) PA

Figure 4.6: Timeline for various epitope scores. Each plot depicts the predicted epitope scores *versus* the sampling year the sequences for which the epitopes scores were calculated, were obtained.

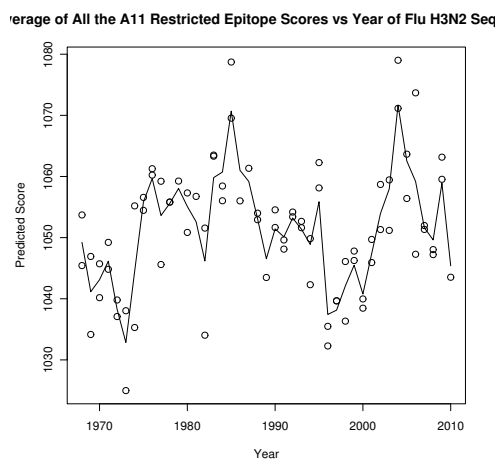
(a) Total Epitope Scores for All Predicted Epitopes (b) Average Epitope Scores for All Predicted Epitopes



(c) Total Epitope Scores for Optimal Epitopes (d) Average Epitope Scores for Optimal Epitopes



(e) Total Epitope Scores for all HLA\*A1101 restricted Epitopes





#### 4.2.4 Clustering Results

##### HIV Epitope Clustering Results

It was mentioned before that clustering of sequences together requires a prior category of clustering. In this case, the levels of HIV in the blood the sequences were obtained from was decided upon. However, as mentioned in Section 4.2.3 on page 100 no direct association could be found between the epitope variants and levels of HIV in the blood. Still, clustering was attempted on the HIV sequences because the author still tested whether HIV can group sequences together based on similar epitopes. The clustering was by and large limited to the HLA\*A0201 restricted optimally defined epitopes. Clustering can be performed in a symmetrical and asymmetrical manner. Especially in the context of “missing epitopes” within certain sequences, the asymmetric method of measuring the distance between two sequences having a different amount of epitopes decreases the distance from the one having the epitope and one that doesn’t. Furthermore, the weights of the epitopes are also taken into account as measure of distance. Epitopes with a low score, if predicted correctly, would have less of an impact on immunogenicity if mutations occur within them. The symmetrical measurement takes the maximum distance between two epitopes of two sequences as the distance measurement for that epitope. The distance for missing epitopes was set to 0.20. The purpose here is to see how well the clustering will perform with high sequence entropy. For this reason, clustering was performed by excluding and including the epitope AIIRILQQL and its variants during clustering. Figure 4.7 on page 113 depicts four heatmaps. The central part of the plot shows the distances between the sequences (marked on the edges) as a colour range from dark blue to white; white being the most distant and dark blue being the closest. Each heatmap is the result of clustering of sequences obtained from patients containing the HLA\*A02xx allotype. Figures 4.7a and 4.7b are the result of clustering with and without epitope AIIRILQQL and its variants. It can be seen that the inclusion of AIIRILQQL had a dramatic influence on introducing noise into the cluster. However, it could still clearly define the two main groups as seen on the heatmaps. Definition of the smaller groups were less clear with the inclusion of AIIRILQQL. The author does not suggest that certain epitopes need be excluded or are invaluable in determining sequences with similar immunogenic patterns, but merely to illustrate the limitation of a sequence-only approach. Figure 4.7c shows clustering performed with the asymmetric method. With symmetric clustering groups are mirrored on either side of the heatmap. With asymmetric clustering, unique groups are defined, meaning that any arbitrary group on the heatmap is not reflected at any other part of the heatmap. Because the differences in weight are also taken account during both measurements (different weights used as a factor of the Frankild distance), lighter regions in the heatmap also indicate a predicted difference in potency between the sets of sequences. The final Figure, 4.7d shows clustering by using all the

predicted 9-mer epitopes of HIV. This heatmap clearly shows that clustering is hampered when the amount of potential epitopes to be considered is high. HIV, being highly mutable, is of course very susceptible to “noisy” clustering in general.

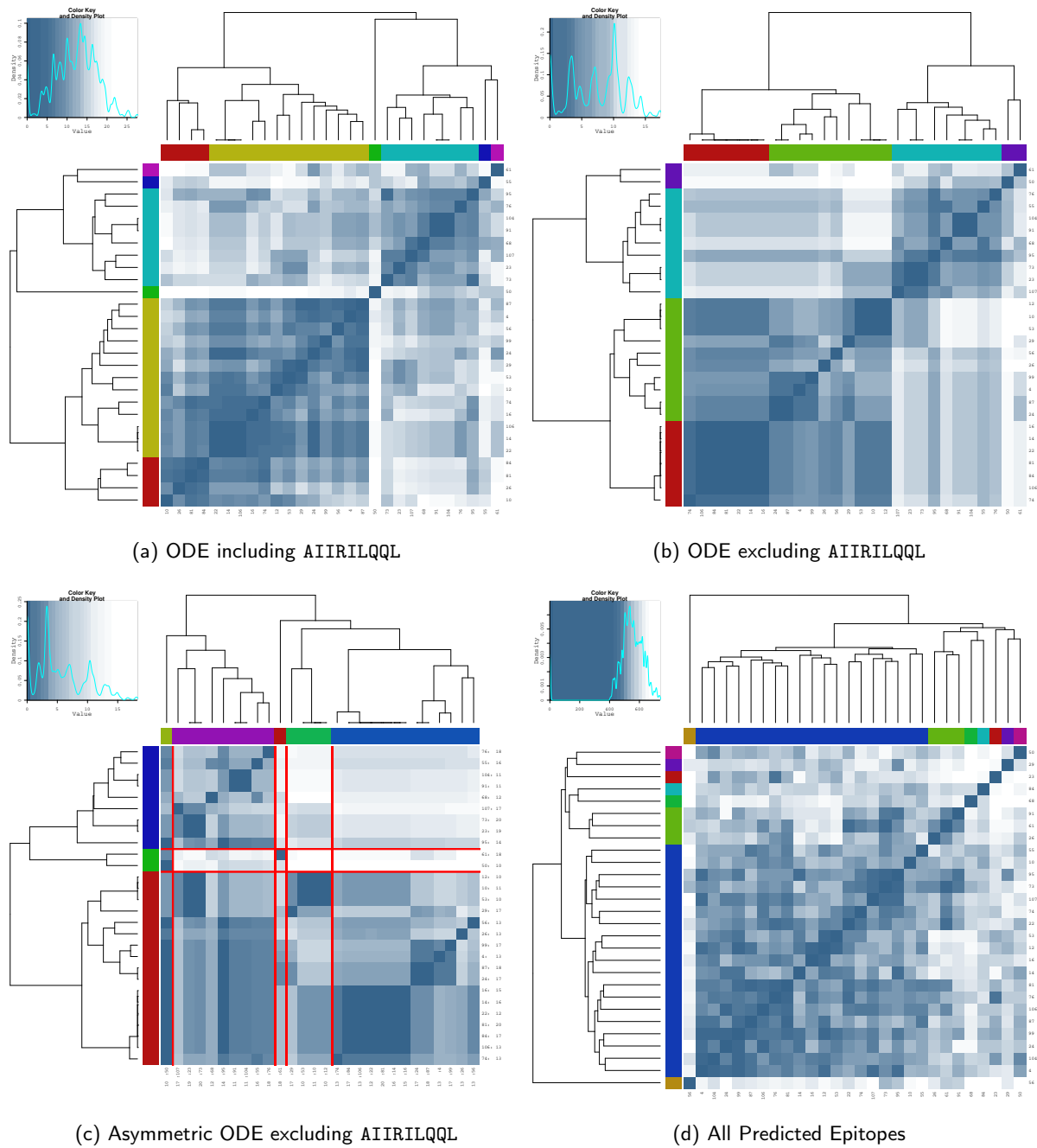


Figure 4.7: Below are the results of clustering performed on sequences from patients with the HLA\*A02xx HLA allotype.

### Influenza Epitope Clustering

Clustering of Influenza epitopes were done in a similar manner to HIV epitopes, with the difference being the attribute chosen as hypothetical grouping criterium. The sequences obtained from LANL all contain at least the year of submission. Converse of the timeline procedure described in the previous section, the epitopes will now be used to cluster dates together. However, first it will be determined if there are differences between the CTL epitope repertoire of Influenza serotypes H1N1 and H3N2. The clustering was done symmetrically. The result of the separation based on CTL epitope repertoire is shown in Figure 4.8 on the following page. For both the optimal and predicted epitopes, there is a clear distinction between H1N1 and H3N2 CTL epitope repertoires (see the description of Figure 4.8. Still, a few sequences from the other serotype are included in both H1N1 and H3N2 groups. This could be the result of recombination between H3N2 and H1N1 genomes during simultaneous infection of both.

**Clustering of H1N1 Sequences** Clustering of H1N1 sequences was performed using only the optimally defined epitopes. The clustering is shown in Figure 4.9 on page 116. On closer inspection, it is clear that the 2009 H1N1 strain's sequences are separate from the rest. To investigate whether this is simply because the H1N1 sequences are naturally grouped together because of their sheer number or whether they are separated from the rest of the H1N1 sequences due to a difference in epitopes, the epitope differences were obtained and are displayed in Table 4.20 on the following page. This is not a complete list of the optimal epitopes, but merely those for which a large difference was observed. As can be seen in the Table, the epitope variant of **RLNKRSYLI**, **RLNKRGYLI** differs by only one amino acid, but that substitution does occur in the central part of the epitope, which has an influence on cross-reactivity and immunogenicity. This epitope is also lowly conserved, having six other variants. Other epitopes with central/near central substitutions are the epitopes **WMMAMRYPI** and **LLMDALKLSI**, although two thirds of the epitope variants for the pre-2009 Influenza H1N1 strains have **LLMALKLSI**. The epitopes **LLRSAIGPV** is an epitope which is predicted to exist in one of the pre-2009 H1N1 sequences. This Table illustrates the power of **Fortuna** to differentiate between two clusters.

**Clustering of H3N2 Sequences** Clustering of H3N2 sequences was performed using the optimal epitopes and all the predicted 9-mer epitopes. The heatmaps are shown in Figure 4.10 on page 117. Here, the attempt was made to isolate sequences obtained during the 1968-69 Hong Kong flu pandemic from the rest of the sequences. The optimal epitopes for Influenza could not isolate sequences of this time period from the rest of the sequences. However, when all the predicted 9-mer epitopes were used for clustering, sequences from and around 1969 were, in fact, isolated from the rest. Although noise is introduced into the clustering when considering a large

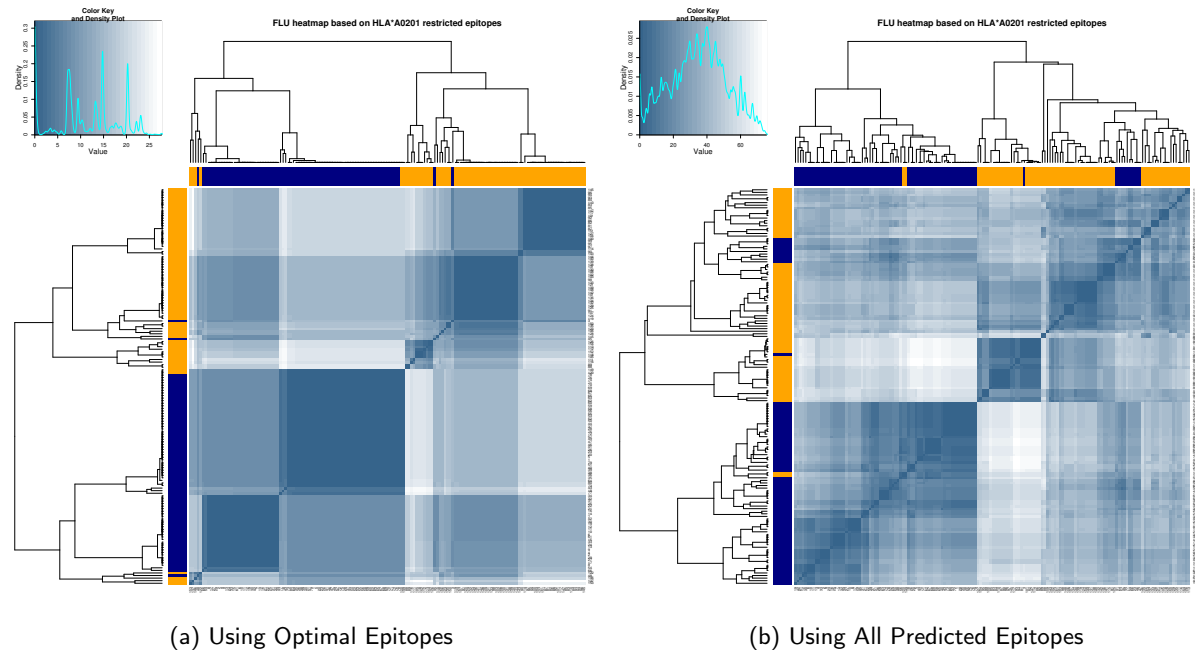


Figure 4.8: The two heatmaps in this figure depict separation of H3N2 and H1N1 serotypes based on their CTL repertoire. The first heatmap shows cluster-based separation on optimally defined epitopes while the second heatmap shows separation based on all the predicted epitopes restricted to HLA\*A0201. In both cases, the output from the pathway prediction results were used as weights. The sides of the heatmap marked orange represent sequences from H1N1 while the blue represents sequences from H3N2.

Table 4.20: CompSeq

| Position | Frequency Diff | Weight Diff | H1N1 2009                       | H1N1 Other   | SeqLogo Group 1 | SeqLogo Group 2 |
|----------|----------------|-------------|---------------------------------|--|-----------------|-----------------|
| 1523     | 0.03           | -0.78       | RLNKRGYLI (42)                  | KLTRRSYLI (1)<br>KLNRRGYLI (1)<br>KLNRRGYLI (1)<br>RLNKRSSYLI (1)<br>RLNKRSSYLI (37)<br>KLDKRSYLI (25)<br>KLNKRSSYLI (5) |                 |                 |
| 2139     | 0.00           | -0.09       | WMMANKYPI (1)<br>WMMANKYPI (41) | WMMANKYPI (72)<br>WMMANKYPI (1)  |                 |                 |
| 3418     | -0.01          | -0.44       |                                 | LLRSAICPV (1)  | None            |                 |
| 3153     | 0.00           | 0.88        | LLMDALKLSI (42)                 | LLMDALKLSI (46)<br>LLMDSLKLSI (27)   |                 |                 |

amount of putative epitopes, it does also reveal subtle differences between the sequences and even with the included “noise” in the comparison, there are still clearly definable clusters.

### 4.2.5 Self-Epitope Discovery

The predicted epitopes of HLA\*A0201 for HIV and Influenza were scanned for self-epitopes using the results of a BLASTP search referencing human proteins. Both HIV and Influenza contain epitopes with high Frankild scores with respect to potential self-epitopes. For influenza, the optimal epitope GILGFVFTL is closely related to the sequence GILLFLFTL although the G->L mutation is in the center region of the supposed self-epitope and would probably result in non-cross reactivity. The highly mutable HIV epitope, AIIRILQQL has a high Frankild score when

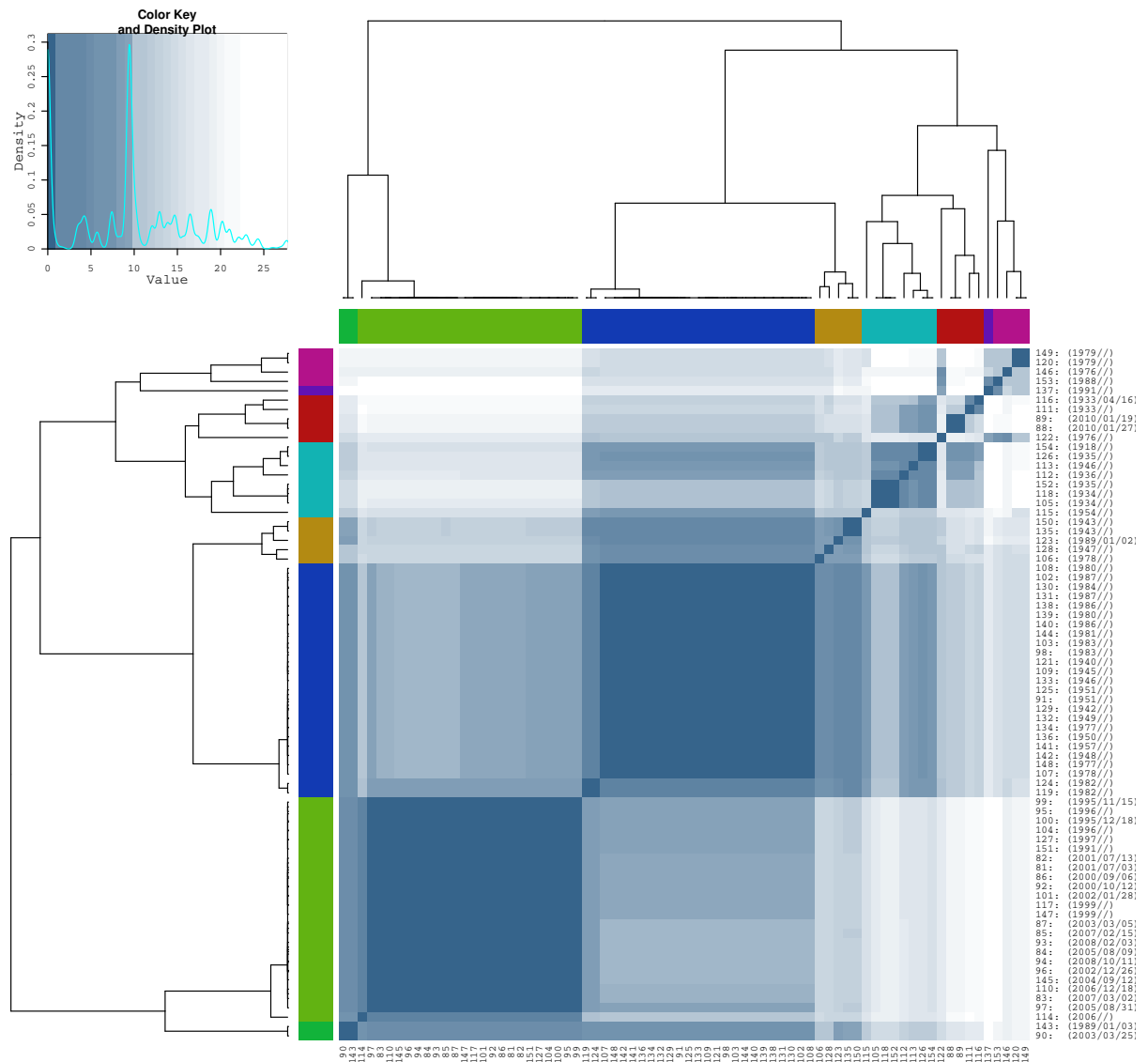
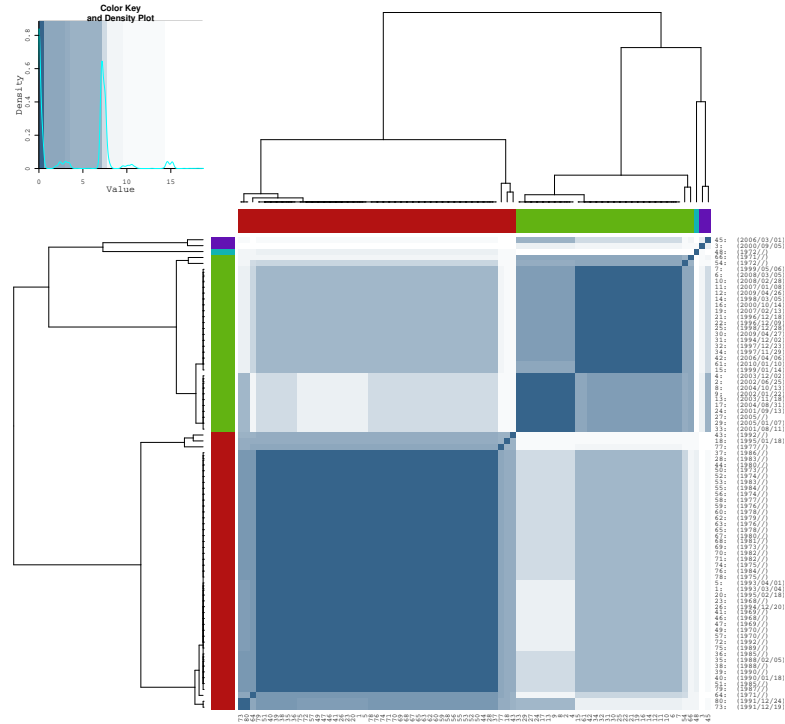
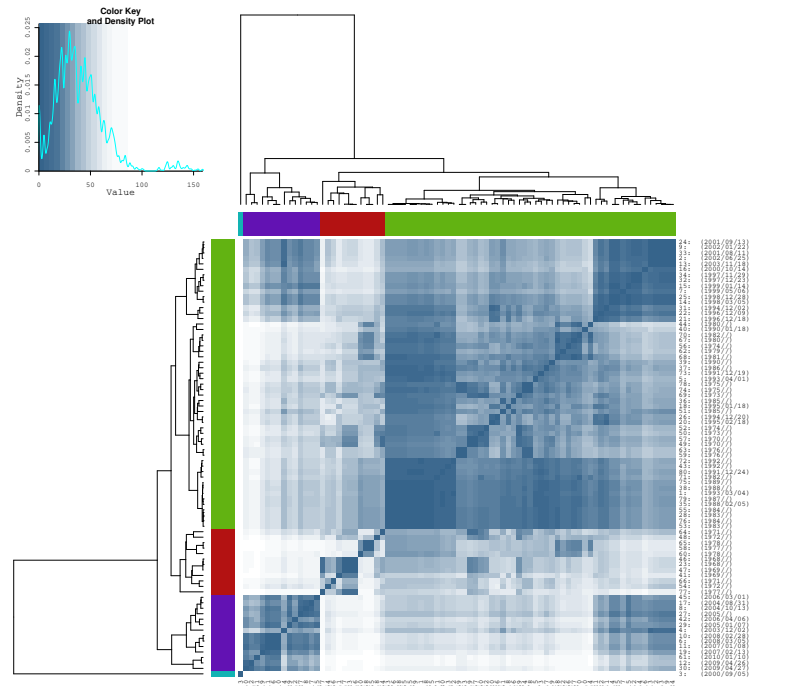


Figure 4.9: The heatmap shows the clustering of all the Influenza A H1N1 sequences, including sequences from the 2009 strain.

compared to AIQRVLQQL which originates from the zinc-finger domain of an untitled protein. It is very likely that many cross-reactivity with self-epitopes were missed by using the Frankild score as noted by the authors of that study. However, it has been shown that closely related matches can still be found. Curiously, by examining Figure 4.11 on page 118 it is evident that a lot of the epitopes detected have high Frankild scores when compared to potential self-epitopes. Although it seems that Influenza epitopes have a higher incidence of self-epitopes, it should be noted that each dot on the plot represents an epitope variant and because HIV is highly mutable, many of the variants of the epitopes will occur in the lower frequency regions.



(a) Optimal Epitopes



(b) All Predicted 9-mer Epitopes

Figure 4.10: The two Figures show the clustering of Influenza A serotype H3N2 according to HLA\*A0201 restricted epitopes. The first figure is generated by using the optimally defined epitopes while the second figure is generated by using all the 9-mer epitopes predicted.

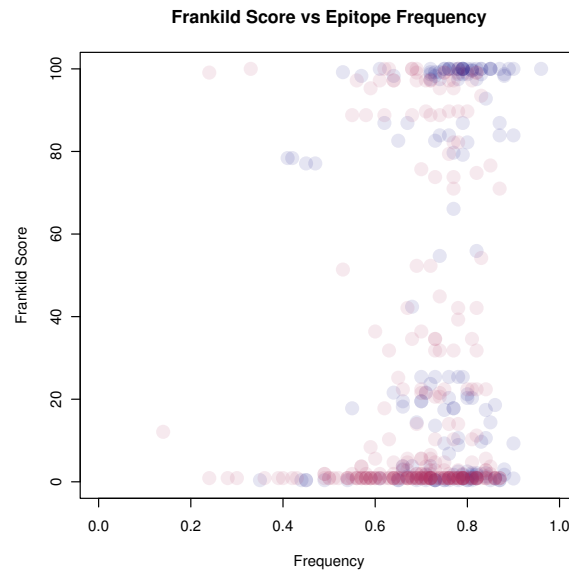


Figure 4.11: This figure depicts the average Frankild score and frequency of predicted HLA\*A0201 restricted epitopes and their variants. Blue represents scores obtained from Influenza epitopes while red represents scores obtained from HIV epitopes.

### 4.3 Conclusion

In this chapter, Fortuna was used to analyse the epitopes of two important pathogens, namely HIV and Influenza. It was shown that although only a modest amount of reliance can be put in using prediction tools to correctly identify immunogenic CTL epitopes, the analyses performed from these predictions did aid in the discovery of potential epitopes. In the case of Influenza, by examining the variants of epitopes predicted and those from the literature, relationships could be established between occurrence of an epitope and incidence of a pandemic. For HIV, there was no correlation between HIV viral load and epitope variant occurrence, but it did show that the variants of the HIV epitope SLYNTVATL is associated with the extreme values of viral load. Although the author could not perform the desired clustering of HIV sequences by HIV level based on CTL epitope repertoire, HIV provided some insight into the limitations of using a direct sequence comparison approach when trying to group epitopes together. For Influenza, the clustering did prove useful in isolating pandemic strains. Further analysis of the differences between said pandemic strains and “moderate” strains revealed insight into a novel CTL epitope repertoire. Whether these differences in epitope sequences contributed directly to the virulence of the 2009 Flu strain can only be validated through experiments.



## Conclusionary Discussion

Herein, the development of a method to analyse the CTL epitope profile of multiple sequences for use in immunological comparisons was discussed. Fortuna is a web-accessible tool that combines the results of different tools associated Proteasomal cleavage, TAP affinity, MHC affinity and immunogenicity prediction in an attempt to give a quantitative score representing the fate of a peptide as it goes through the Class I restricted antigen presentation pathway. A novel predictor for one of the steps in the pathway, namely TAP ligand affinity, was created and dubbed Variable Lengthed TAP Predictor (VLTAPP). VLTAPP was shown to perform comparably with other known predictors of TAP affinity. It was also shown how the training of a predictor could provide insight into the the binding region of a molecule for which the structure is not fully known. By analyzing the CTL epitopes of two pathogens, HIV and Influenza, the virtues as well as the shortcomings of prediction tools in the context of CTL epitope prediction was illustrated through epitope score prediction, epitope variant summary and cluster analyses.

### 5.1 Identification of Pitfalls in CTL Epitope Prediction

Throughout the project, the author noted some very obvious and somewhat less obvious pitfalls associated with CTL epitope prediction. A full discussion of each goes beyond the scope of this project, and these shortcomings will only be briefly discussed.

#### 5.1.1 Problems with POPI and Immunogenicity Prediction in General

Of all the predicted steps, immunogenicity has the largest impact on whether a peptide-MHC complex would be recognised by the immune system. The author noted that many epitopes undergo mutations that do not affect the ability to bind to MHC, but change the parts of the epitope that interface with the TCR. The immunogenicity predictor used in this study, POPI, has relatively good validation scores with approximately 60% accuracy of predicting any one of four classes. The author trained POPI as instructed in the literature and after validation, similar

accuracies were obtained. However, the very way in which POPI does predictions is flawed. By circumventing the issue of length of an MHC ligand when calculating immunogenicity, the authors of POPI decided to average all the pre-selected amino acid properties over the entire length of the sequence. This means that any permutation of any sequence would produce the same immunogenicity score. It is very unlikely that  $9! = 362880$  permutations of one epitope will produce the same immunogenicity. The author appreciates that the training set used during construction of POPI is the only real one of its kind, outdated and provides experimental results as discrete values. In a perfect scenario, an immunogenicity predictor would consider the peptide-MHC complex and search for the appropriate T-Cell Receptor  $\alpha\beta$  chains that would have high affinity to the complex. This is a non-trivial task, since it requires prior modelling of millions of  $\text{TCR}_{\alpha\beta}$  combinations in conjunction with correctly docking an arbitrary protein to an arbitrary HLA allotype. Any tool developed that could make these accurate predictions would take a fairly long time to run and would defeat the purpose here; to provide an easily accessible tool to aid in crude possibly pre-experimental CTL epitope analysis.

The other problem with predicting immunogenicity is the differences between the T-Cell repertoire between individuals. Since naïve T-Cells undergo both positive and negative selection depending on what is presented to them in the thymus, there should be a difference in repertoire between people having different sets of HLA molecules. HLA molecules, even within the same supertype have a vastly different set of peptides they present and as a consequence, the naïve T-Cell population would receive vastly different “training”. So, in order to properly predict immunogenicity, it should first be predicted what possible T-Cell clones would be available to scrutinise the peptide-MHC complex based on the molecules presented by the individual’s HLA set. The author cannot imagine a quick solution to the problem with the current knowledge of TCR-peptide-MHC interaction.

Immunogenicity could also be abrogated by indirect factors. In a recent study, it was determined that the HLA\*B2705 restricted epitope, KRWIIILGLNK undergoes a L6  $\rightarrow$  M mutation that attenuates the immune response in general (Lichterfeld *et al.*, 2007). After investigation, it was determined that KRWIIIMLGLNK is still highly immunogenic, but cross reactive with the ILT4 expressed on myelomonocytic cells. Binding of a and MHC-peptide complex to this molecule induces a tolerance response by affecting the maturation of dendritic cells, which are regulators and initiators of the adaptive immune responses. This is an example of a mutation whose effect is not limited to the epitope itself, but can affect responses against multiple antigenic stimuli. These tolerance inducing peptides were also discovered by the researchers in Epstein-Barr Virus (EBV) and Hepatitis C Virus (HCV) showing that this escape mechanism is not all that uncommon.

### 5.1.2 Cross-reactivity

It was only recently that Frankild *et al.* (2008) tackled the issue of CTL epitope cross-reactivity. They developed a simple and modestly effective method to identify cross-reactive CTL epitopes. They did put a lot of emphasis on CTL epitopes that are cross reactive but only share one or two amino acids. Cross-reactivity can be defined as the ability of two epitopes eliciting a similarly strong CTL immune response by stimulating a similar set of CTL clones. CTL epitopes are not acted upon by a single clone. Therefore, correct “prediction” of cross-reactivity would be predicting the  $\text{TCR}_{\alpha\beta}$  chains that will bind each of the two epitopes and evaluating “shared T-Cell clones” of the epitopes.

### 5.1.3 Proteasomal Cleavage Prediction

The way proteasomal cleavage is handled here is rather simplistic. It utilises a method based on a position specific scoring matrix when it is known that proteasomal cleavage is far more complicated than merely “motive” driven (Piwko and Jentsch, 2006). Something that further compounds the problem is the lack of available proteasomal cleavage sets and as shown with the evaluation of VLTAPP, training set size does have a profound impact on predictor performance.

### 5.1.4 MHC Prediction

Although generally very good, MHC prediction is also subject to the availability of a predictor for the desired HLA allotype. HLA ligand sets are not equal in size for a lot of HLA allotypes with some containing less than a hundred. MHC, being one of the most polymorphic molecules, means that producing training sets for all of the thousands of HLA allotypes known is unfeasible. There have been attempts to try and make predictions for an HLA supertype, but subtle differences between HLA allotypes within the same supertype could also have a profound effect on certain peptide ligands (Alexander *et al.*, 2010). Recently an attempt was made to make predictions for 9-mer MHC ligands using any arbitrary HLA allotype. The way the prediction works is to train a neural network that takes into account the amino acid residues in the binding pockets of the MHC molecule (Lundegaard *et al.*, 2008). So, if a novel HLA allotype is found, approximations could be made for ligand affinity to it as long as the amino acids from the binding pockets are known.

## 5.2 Bioinformatics Facilitating CTL Based Vaccine Design

The concept of using computational tools in aiding the design of vaccines has been reviewed multiple times in the literature (Groot and Berzofsky, 2004, Groot, 2006, Korber *et al.*, 2006). With the increasing availability of efficient sequencing technologies, the genomes of many organisms

can be readily determined and from that, protein sequences. By using MHC prediction alone, a rough set of MHC ligands, and by extension potential CTL epitopes can be elucidated from the sequence alone. The total length of the Influenza proteins used in this study is approximately 3000 amino acids long. To determine affinity to a single HLA molecule by experimentation alone, taking into account peptide lengths of 8-11 amino acids would be considered, would require approximately 12000 assays without counting replication. If an MHC predictor predicted 300 MHC ligands for each of the peptide lengths, this would reduce the amount of experimentation needed by 90%. Either this would allow more modest budgets to perform the experiments or more replicates of the experiments to be performed, increasing the result of the experiments. By using the sequences of mutational variants of the same protein, peptides can be filtered further based on their sequence conservation. A moderately immunogenic peptide with high sequence conservation would make a more appropriate candidate for a vaccine than one with high immunogenicity but with low sequence conservation, since the pathogen of origin is likely to quickly acquire escape mutations within this region.

### 5.3 The use of Prediction Tools in this Study

Using prediction tools outright for inferring immunological relationships between protein variants is bold to say the least, but between all the incorrect predictions, the author did discover a very significant hint of correct classification. Indeed, by including only the optimal epitopes, classification was by and large improved, with the exception of isolating the Influenza H3N2 strain of the Hong Kong pandemic. By knowing virulent strains of Influenza and CTL epitope variants associated with them, it could be possible to estimate whether current trends in Influenza mutations might lead to reversion of a strain to its pandemic state. It is very possible to use simple phylogenetic methods to make these predictions, however the author notes that applying phylogenetic methods would take the mutations associated with all selective pressures into account and not necessarily isolate appropriate mutations associated with immune surveillance escape. The author suspects that if the prediction methods are more accurate and more appropriate weight assignment can be achieved, the total set of predicted epitopes should be sufficient in detecting immunological similarities, since non-immunogenic epitopes would be down-weighted enough for changes in them to become of no consequence when comparing two or more CTL epitope profiles.

### 5.4 Conclusion

With the increase in size of available immunological datasets the author hopes that the methods used in this project to aid in immunological classification can be improved. With high throughput

accurate computational analysis of immunological profiles it may be possible in future to start designing tailor made CTL epitope based vaccines. That is, vaccines specifically designed for the HLA Allotype profile of the recipient. Especially in the context where treatment is not readily available, vaccines can severely reduce the mortality rate in regions afflicted by vaccine preventable disease.

The project presented here, aims to be a stepping stone in the future development of computational immunology tools. Presented to the reader were the advances made in recent years in applying computational methods to the field of Immunology. The reader was also exposed to the rationale behind creating predictors and how they can aid in research. Development of Fortuna allows to analyse the CTL epitope profiles of different proteins, compare them with each other and evaluate similar “immune profile”. Further development of Fortuna would include fine-tuning MHC prediction, especially in the context of correctly predicting affinity values for and MHC ligands with respect to its variants, better prediction of immunogenicity and better abstraction of cross-reactivity measurements. It should be interesting to see the application of Fortuna or similar method to other diseases where CTL-epitopes could play a protective role, for example most cancers and parasitic infections such as visceral Leishmaniasis. Also, in the study of finding close matches to avoid host and graft immunological incompatibility after organ transplants. Having said that, there is still a lot of research necessary to assimilate necessary data to make more accurate predictions. Especially in the context of pMHC-TCR interaction that is, despite the extreme efforts put in by researchers, still poorly understood. Bioinformatics tools in general, also have a tendency to make people complacent with prediction/calculation results. It should always be remembered that prediction tools perform as best as they are named and can be, almost in a fashion of fate, completely wrong when all supposed evidence suggest they are correct.

## Summary

The field of Immunology can be hindered by labour intensive experimental procedures. In the context of CTL epitope studies, computational methods can aid in reducing the redundancy of elucidating epitopes contained within protein sequences by revealing peptides that are likely to be epitopes. By combining the prediction results of proteasomal cleavage, TAP affinity, MHC affinity and Immunogenicity putative epitopes can be revealed. Extending this procedure is done by performing the analysis on multiple sequences to reveal plausible escape mutations through visualising all the variants of a potential epitopes. Cluster analysis can reveal how different sequences group according to their epitope profile. By integrating this functionality into a web-based application, researchers investigating the CTL epitopes of arbitrary proteins are aided in the analysis and discovery of potential epitopes.

## Bibliography

- Alexander, J., Bilsel, P., del Guercio, M.-F., Marinkovic-Petrovic, A., Southwood, S., Stewart, S., Ishioka, G., Kotturi, M. F., Botten, J., Sidney, J., Newman, M., and Sette, A. (2010) Identification of broad binding class i hla supertype epitopes to provide universal coverage of influenza a virus. *Hum Immunol*, **71** (5), 468–474.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J Mol Biol*, **215** (3), 403–410.
- Barlow, R. B., Bond, S. M., Bream, E., Macfarlane, L., and McQueen, D. S. (1997) Antagonist inhibition curves and the measurement of dissociation constants. *Br J Pharmacol*, **120** (1), 13–18.
- Bhasin, M. and Raghava, G. P. S. (2004) Analysis and prediction of affinity of tap binding peptides using cascade svm. *Protein Sci*, **13** (3), 596–607.
- Bhasin, M., Singh, H., and Raghava, G. P. S. (2003) Mhcdbn: a comprehensive database of mhc binding and non-binding peptides. *Bioinformatics*, **19** (5), 665–666.
- Brusic, V., Bajic, V. B., and Petrovsky, N. (2004) Computational methods for prediction of t-cell epitopes—a framework for modelling, testing, and applications. *Methods*, **34** (4), 436–443.
- Brusic, V., Rudy, G., and Harrison, L. C. (1998) Mhcpep, a database of mhc-binding peptides: update 1997. *Nucleic Acids Res*, **26** (1), 368–371.
- Buus, S., Lauemøller, S. L., Worning, P., Kesmir, C., Frimurer, T., Corbet, S., Fomsgaard, A., Hilden, J., Holm, A., and Brunak, S. (2003) Sensitive quantitative predictions of peptide-mhc binding by a 'query by committee' artificial neural network approach. *Tissue Antigens*, **62** (5), 378–384.
- Callahan, M. K., Garg, M., and Srivastava, P. K. (2008) Heat-shock protein 90 associates with n-terminal extended peptides and is required for direct and indirect antigen presentation. *Proc Natl Acad Sci U S A*, **105** (5), 1662–1667.
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L. (2009) Blast+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
- Carugo, O. (2007) Detailed estimation of bioinformatics prediction reliability through the fragmented prediction performance plots. *BMC Bioinformatics*, **8**, 380.
- Chang, S.-C., Momburg, F., Bhutani, N., and Goldberg, A. L. (2005) The er aminopeptidase, erap1, trims precursors to lengths of mhc class i peptides by a "molecular ruler" mechanism. *Proc Natl Acad Sci U S A*, **102** (47), 17107–17112.

- Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M. J. L. (2009) Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25** (11), 1422–1423.
- Cresswell, P., Ackerman, A. L., Giodini, A., Peaper, D. R., and Wearsch, P. A. (2005) Mechanisms of mhc class i-restricted antigen processing and cross-presentation. *Immunol Rev*, **207**, 145–157.
- Crooks, G. E., Hon, G., Chandonia, J.-M., and Brenner, S. E. (2004) Weblogo: a sequence logo generator. *Genome Res*, **14** (6), 1188–1190.
- Daniel, S., Brusica, V., Caillat-Zucman, S., Petrovsky, N., Harrison, L., Riganelli, D., Sinigaglia, F., Gallazzi, F., Hammer, J., and van Endert, P. M. (1998) Relationship between peptide selectivities of human transporters associated with antigen processing and hla class i molecules. *J Immunol*, **161** (2), 617–624.
- Dunn, J. C. (1973) A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, **3** (3), 32–57.
- Dönnes, P. and Elofsson, A. (2002) Prediction of mhc class i binding peptides, using svmhc. *BMC Bioinformatics*, **3**, 25–32.
- Dönnes, P. and Kohlbacher, O. (2006) Svmhc: a server for prediction of mhc-binding peptides. *Nucleic Acids Res*, **34** (Web Server issue), W194–W197.
- Eleuteri, A. M., Kohanski, R. A., Cardozo, C., and Orłowski, M. (1997) Bovine spleen multi-catalytic proteinase complex (proteasome). replacement of x, y, and z subunits by lmp7, lmp2, and mecl1 and changes in properties and specificity. *J Biol Chem*, **272** (18), 11824–11831.
- Fahnestock, M. L., Johnson, J. L., Feldman, R. M., Tsomides, T. J., Mayer, J., Narhi, L. O., and Bjorkman, P. J. (1994) Effects of peptide length and composition on binding to an empty class i mhc heterodimer. *Biochemistry*, **33** (26), 8149–8158.
- Frankild, S., de Boer, R. J., Lund, O., Nielsen, M., and Kesmir, C. (2008) Amino acid similarity accounts for t cell cross-reactivity and for "holes" in the t cell repertoire. *PLoS One*, **3** (3), e1831.
- Gaczynska, M., Goldberg, A. L., Tanaka, K., Hendil, K. B., and Rock, K. L. (1996) Proteasome subunits x and y alter peptidase activities in opposite ways to the interferon-gamma-induced subunits lmp2 and lmp7. *J Biol Chem*, **271** (29), 17275–17280.
- Ginodi, I., Vider-Shalit, T., Tsaban, L., and Louzoun, Y. (2008) Precise score for the prediction of peptides cleaved by the proteasome. *Bioinformatics*, **24** (4), 477–483.
- Groot, A. S. D. (2006) Immunomics: discovering new targets for vaccines and therapeutics. *Drug Discov Today*, **11** (5-6), 203–209.
- Groot, A. S. D. and Berzofsky, J. A. (2004) From genome to vaccine—new immunoinformatics tools for vaccine design. *Methods*, **34** (4), 425–428.
- Hakenberg, J., Nussbaum, A. K., Schild, H., Rammensee, H.-G., Kuttler, C., Holzhütter, H.-G., Kloetzel, P.-M., Kaufmann, S. H. E., and Mollenkopf, H.-J. (2003) Mappp: Mhc class i antigenic peptide processing prediction. *Appl Bioinformatics*, **2** (3), 155–158.
- Henikoff, S. and Henikoff, J. G. (1992) Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, **89** (22), 10915–10919.
- Hennig, C. (2009) In *fpc: Fixed point clusters, clusterwise regression and discriminant plots*. R package version 1.2-7.



- Holzhütter, H. G., Frömmel, C., and Kloetzel, P. M. (1999) A theoretical approach towards the identification of cleavage-determining amino acid motifs of the 20 s proteasome. *J Mol Biol*, **286** (4), 1251–1265.
- Iversen, A. K. N., Stewart-Jones, G., Learn, G. H., Christie, N., Sylvester-Hviid, C., Armitage, A. E., Kaul, R., Beattie, T., Lee, J. K., Li, Y., Chotiyarnwong, P., Dong, T., Xu, X., Luscher, M. A., MacDonald, K., Ullum, H., Klarlund-Pedersen, B., Skinhøj, P., Fugger, L., Buus, S., Mullins, J. I., Jones, E. Y., van der Merwe, P. A., and McMichael, A. J. (2006) Conflicting selective forces affect t cell receptor contacts in an immunodominant human immunodeficiency virus epitope. *Nat Immunol*, **7** (2), 179–189.
- Jim Lemon, Ben Bolker, S. O., Eduardo Klein, B. R., Hadley Wickham, A. T., Olivier Eterradosi, G. G., Michael Toews, John Kane, M. C., Rolf Turner, Carl Witthoft, J. S., Thomas Petzoldt, Remko Duursma, E. B., and Levy, O. (2009) In *plotrix: Various plotting functions*. R package version 2.6-4.
- Kesmir, C., Nussbaum, A. K., Schild, H., Detours, V., and Brunak, S. (2002) Prediction of proteasome cleavage motifs by neural networks. *Protein Eng*, **15** (4), 287–296.
- Khan, A. R., Baker, B. M., Ghosh, P., Biddison, W. E., and Wiley, D. C. (2000) The structure and stability of an hla-a\*0201/octameric tax peptide complex with an empty conserved peptide-n-terminal binding site. *J Immunol*, **164** (12), 6398–6405.
- Kisselev, A. F., Akopian, T. N., Woo, K. M., and Goldberg, A. L. (1999) The sizes of peptides generated from protein by mammalian 26 and 20 s proteasomes. implications for understanding the degradative mechanism and antigen presentation. *J Biol Chem*, **274** (6), 3363–3371.
- Korber, B., LaBute, M., and Yusim, K. (2006) Immunoinformatics comes of age. *PLoS Comput Biol*, **2** (6), e71.
- Korber, B. T. M., Brander, C., Haynes, B. F., Koup, R., Moore, J. P., Walker, B. D., , and Watkins, D. I., editors (2007) In *HIV Molecular Immunology 2006/2007*. Los Alamos National Laboratory, Theoretical Biology and Biophysics, Los Alamos, New Mexico.
- Kyte, J. and Doolittle, R. F. (1982) A simple method for displaying the hydropathic character of a protein. *J Mol Biol*, **157** (1), 105–132.
- Larsen, J. E. P., Lund, O., and Nielsen, M. (2006) Improved method for predicting linear b-cell epitopes. *Immunome Res*, **2**, 2.
- Larsen, M. V., Lundegaard, C., Lamberth, K., Buus, S., Brunak, S., Lund, O., and Nielsen, M. (2005) 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*, **35** (8), 2295–2303.
- Lefranc, M.-P. (2005) Imgt, the international immunogenetics information system: a standardized approach for immunogenetics and immunoinformatics. *Immunome Res*, **1**, 3.
- Lichterfeld, M., Kavanagh, D. G., Williams, K. L., Moza, B., Mui, S. K., Miura, T., Sivamurthy, R., Allgaier, R., Pereyra, F., Trocha, A., Feeney, M., Gandhi, R. T., Rosenberg, E. S., Altfeld, M., Allen, T. M., Allen, R., Walker, B. D., Sundberg, E. J., and Yu, X. G. (2007) A viral ctl escape mutation leading to immunoglobulin-like transcript 4-mediated functional inhibition of myelomonocytic cells. *J Exp Med*, **204** (12), 2813–2824.
- Limas, M. C., Meré, J. B. O., González, E. P. V., de Pisón Ascacibar, F. J. M., Espinoza, A. V. P., and Elías, F. A. (2007) In *AMORE: A MORE flexible neural network package*.

- Lin, H. H., Ray, S., Tongchusak, S., Reinherz, E. L., and Brusic, V. (2008) Evaluation of mhc class i peptide binding prediction servers: applications for vaccine research. *BMC Immunol*, **9**, 8.
- Lundegaard, C., Lund, O., and Nielsen, M. (2008) Accurate approximation method for prediction of class i mhc affinities for peptides of length 8, 10 and 11 using prediction tools trained on 9mers. *Bioinformatics*, **24** (11), 1397–1398.
- Martin, O. V., Lai, K. M., Scrimshaw, M. D., and Lester, J. N. (2005) Receiver operating characteristic analysis for environmental diagnosis. a potential application to endocrine disruptor screening: in vitro estrogenicity bioassays. *Environ Sci Technol*, **39** (14), 5349–5355.
- Mason, D. (1998) A very high level of crossreactivity is an essential feature of the t-cell receptor. *Immunology Today*, **19** (9), 395 – 404.
- Matthews, B. W. (1975) Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochim Biophys Acta*, **405** (2), 442–451.
- Nielsen, M., Lundegaard, C., Lund, O., and Kesmir, C. (2005) The role of the proteasome in generating cytotoxic t-cell epitopes: insights obtained from improved predictions of proteasomal cleavage. *Immunogenetics*, **57** (1-2), 33–41.
- Nielsen, M., Lundegaard, C., Worning, P., Hvid, C. S., Lamberth, K., Buus, S., Brunak, S., and Lund, O. (2004a) Improved prediction of mhc class i and class ii epitopes using a novel gibbs sampling approach. *Bioinformatics*, **20** (9), 1388–1397.
- Nielsen, M., Lundegaard, C., Worning, P., Hvid, C. S., Lamberth, K., Buus, S., Brunak, S., and Lund, O. (2004b) Improved prediction of mhc class i and class ii epitopes using a novel gibbs sampling approach. *Bioinformatics*, **20** (9), 1388–1397.
- Ochoa-Garay, J., McKinney, D. M., Kochounian, H. H., and McMillan, M. (1997) The ability of peptides to induce cytotoxic t cells in vitro does not strongly correlate with their affinity for the h-2ld molecule: implications for vaccine design and immunotherapy. *Mol Immunol*, **34** (3), 273–281.
- Pamer, E. and Cresswell, P. (1998) Mechanisms of mhc class i–restricted antigen processing. *Annu Rev Immunol*, **16**, 323–358.
- Parker, K. C., Bednarek, M. A., and Coligan, J. E. (1994) Scheme for ranking potential hla-a2 binding peptides based on independent binding of individual peptide side-chains. *Journal of Immunology*, **152**, 164–175.
- Peters, B., Bui, H.-H., Frankild, S., Nielson, M., Lundegaard, C., Kostem, E., Basch, D., Lamberth, K., Harndahl, M., Fleri, W., Wilson, S. S., Sidney, J., Lund, O., Buus, S., and Sette, A. (2006) A community resource benchmarking predictions of peptide binding to mhc-i molecules. *PLoS Comput Biol*, **2** (6), e65.
- Peters, B., Bulik, S., Tampe, R., Endert, P. M. V., and Holzhütter, H.-G. (2003) Identifying mhc class i epitopes by predicting the tap transport efficiency of epitope precursors. *J Immunol*, **171** (4), 1741–1749.
- Peters, B., Janek, K., Kuckelkorn, U., and Holzhütter, H.-G. (2002) Assessment of proteasomal cleavage probabilities from kinetic analysis of time-dependent product formation. *J Mol Biol*, **318** (3), 847–862.
- Peters, B. and Sette, A. (2005) Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method. *BMC Bioinformatics*, **6**, 132.

- Piwko, W. and Jentsch, S. (2006) Proteasome-mediated protein processing by bidirectional degradation initiated from an internal site. *Nat Struct Mol Biol*, **13** (8), 691–697.
- Pruitt, K. D., Tatusova, T., Klimke, W., and Maglott, D. R. (2009) Ncbi reference sequences: current status, policy and new initiatives. *Nucleic Acids Res*, **37** (Database issue), D32–D36.
- Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2005) Ncbi reference sequence (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, **33** (Database issue), D501–D504.
- Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2007) Ncbi reference sequences (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, **35** (Database issue), D61–D65.
- R Development Core Team (2009) In *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- Ramm, M., Dangoor, K., and Sayfan, G. (2006) In *Rapid Web Applications with TurboGears: Using Python to Create Ajax-Powered Sites (Prentice Hall Open Source Software Development Series)*. Prentice Hall PTR, Upper Saddle River, NJ, USA.
- Rammensee, H.-G., Bachmann, J., Philipp, N., Emmerich, N., Bachor, O. A., and Stevanovic, S. (1999) Syfpeithi: database for mhc ligands and peptide motifs. *Immunogenetics*, **50**, 213–219.
- Rammensee, H. G., Friede, T., and Stevanovic, S. (1995) Mhc ligands and peptide motifs: first listing. *Immunogenetics*, **41** (4), 178–228.
- Reche, P. A. and Reinherz, E. L. (2005) Pepvac: a web server for multi-epitope vaccine development based on the prediction of supertypic mhc ligands. *Nucleic Acids Research*, **33**, Web Server Issue.
- Reche, P. A., Zhang, H., Glutting, J.-P., and Reinherz, E. L. (2005) Epimhc: a curated database of mhc-binding peptides for customized computational vaccinology. *Bioinformatics*, **21** (9), 2140–2141.
- Shannon, C. E. (1948) A mathematical theory of communication, part i. *Bell Syst Tech J*, **27**, 379–423.
- Shedlock, D. J. and Shen, H. (2003) Requirement for cd4 t cell help in generating functional cd8 t cell memory. *Science*, **300** (5617), 337–339.
- Shevach, E. M. (2002) Cd4+ cd25+ suppressor t cells: more questions than answers. *Nat Rev Immunol*, **2** (6), 389–400.
- Sidney, J., Peters, B., Frahm, N., Brander, C., and Sette, A. (2008) Hla class i supertypes: a revised and updated classification. *BMC Immunol*, **9**, 1.
- Sidney, J., Southwood, S., Oseroff, C., del Guercio, M. F., Sette, A., and Grey, H. M. (2001) Measurement of mhc/peptide interactions by gel filtration. *Curr Protoc Immunol*, **Chapter 18**, Unit 18.3.
- Sijts, A. J., Pilip, I., and Pamer, E. G. (1997) The listeria monocytogenes-secreted p60 protein is an n-end rule substrate in the cytosol of infected cells. implications for major histocompatibility complex class i antigen processing of bacterial proteins. *J Biol Chem*, **272** (31), 19261–19268.
- Snyder, H. L., Yewdell, J. W., and Bennink, J. R. (1994) Trimming of antigenic peptides in an early secretory compartment. *J Exp Med*, **180** (6), 2389–2394.
- Stothard, P. (2000) The sequence manipulation suite: Javascript programs for analyzing and formatting protein and dna sequences. *Biotechniques*, **28** (6), 1102, 1104.

- Sylvester-Hvid, C., Kristensen, N., Blicher, T., Ferré, H., Lauemøller, S. L., Wolf, X. A., Lamberth, K., Nissen, M. H., Pedersen, L. Ø., and Buus, S. (2002) Establishment of a quantitative elisa capable of determining peptide - mhc class i interaction. *Tissue Antigens*, **59** (4), 251–258.
- Tenzer, S., Peters, B., Bulik, S., Schoor, O., Lemmel, C., Schatz, M. M., Kloetzel, P.-M., Rammensee, H.-G., Schild, H., and Holzhütter, H.-G. (2005) Modeling the mhc class i pathway by combining predictions of proteasomal cleavage, tap transport and mhc class i binding. *Cell Mol Life Sci*, **62** (9), 1025–1037.
- Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, **22** (22), 4673–4680.
- Toes, R. E., Nussbaum, A. K., Degermann, S., Schirle, M., Emmerich, N. P., Kraft, M., Laplace, C., Zwinderman, A., Dick, T. P., Müller, J., Schönfisch, B., Schmid, C., Fehling, H. J., Stevanovic, S., Rammensee, H. G., and Schild, H. (2001) Discrete cleavage motifs of constitutive and immunoproteasomes revealed by quantitative analysis of cleavage products. *J Exp Med*, **194** (1), 1–12.
- Toseland, C. P., Clayton, D. J., McSparron, H., Hemsley, S. L., Blythe, M. J., Paine, K., Doytchinova, I. A., Guan, P., Hattotuwigama, C. K., and Flower, D. R. (2005) Antigen: a quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical, and cellular data. *Immunome Res*, **1** (1), 4.
- Trost, B., Bickis, M., and Kusalik, A. (2007) Strength in numbers: achieving greater accuracy in mhc-i binding prediction by combining the results from multiple prediction tools. *Immunome Res*, **3**, 5.
- Tung, C.-W. and Ho, S.-Y. (2007) Popi: predicting immunogenicity of mhc class i binding peptides by mining informative physicochemical properties. *Bioinformatics*, **23** (8), 942–949.
- Uebel, S., Kraas, W., Kienle, S., Wiesmüller, K. H., Jung, G., and Tampé, R. (1997) Recognition principle of the tap transporter disclosed by combinatorial peptide libraries. *Proc Natl Acad Sci U S A*, **94** (17), 8976–8981.
- Unno, M., Mizushima, T., Morimoto, Y., Tomisugi, Y., Tanaka, K., Yasuoka, N., and Tsukihara, T. (2002) The structure of the mammalian 20s proteasome at 2.75 a resolution. *Structure*, **10** (5), 609–618.
- Urbanek, S. and Horner, J. (2009) In *Cairo: R graphics device using cairo graphics library for creating high-quality bitmap (PNG, JPEG, TIFF), vector (PDF, SVG, PostScript) and display (X11 and Win32) output..* R package version 1.4-5.
- van Endert, P. M., Riganelli, D., Greco, G., Fleischhauer, K., Sidney, J., Sette, A., and Bach, J. F. (1995) The peptide-binding motif for the human transporter associated with antigen processing. *J Exp Med*, **182** (6), 1883–1895.
- Warnes, G. R. (2009) In *gplots: Various R programming tools for plotting data.* R package version 2.7.1.
- Wright, C. A., Kozik, P., Zacharias, M., and Springer, S. (2004) Tapasin and other chaperones: models of the mhc class i loading complex. *Biol Chem*, **385** (9), 763–778.
- Yusim, K., Kesmir, C., Gaschen, B., Addo, M. M., Altfeld, M., Brunak, S., Chigaev, A., Detours, V., and Korber, B. T. (2002) Clustering patterns of cytotoxic t-lymphocyte epitopes in human immunodeficiency virus type 1 (hiv-1) proteins reveal imprints of immune evasion on hiv-1 global variation. *J Virol*, **76** (17), 8757–8768.

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**, 3.