Chapter 1

Introduction

1.1 Introduction

1.1.1 Overview

Major histocompatibility complexes (MHC) play an integral role in the response of the immune system. It is the complex that is responsible for recognizing antigenic peptides and presenting them for the next chain of immune response. There are two major types of MHC; MHC Class I and Class II (Janeway *et al.*, 2001). These two classes involve processing and presenting antigens to T lymphocytes but differ in the immune response pathway. Whilst MHC Class I presents the antigen for recognition of the cytotoxic T cell, the MHC Class II presents the antigen for recognition of the helper T cell. Studying the peptides that bind to MHC Class II molecules can facilitate the pipeline for vaccine design, reducing the number of identifying helper T cell epitopes (Nielsen *et al.*, 2007). The human counterpart for the MHC Class II is the human leukocyte antigen (HLA) Class II.

The lengths of MHC Class II binding peptides were known to vary, and several publications report different ranges. Kato *et al.* (2003) reported the range as varying from eleven to thirty residues, whilst Lafuente & Reche (2009) published the range as being from nine to twenty-two residues. This quality of variable length complicates the process of building a model that best describes this particular class of binding peptides. According to structural studies, only a core of nine residues would fit into the MHC binding groove, this being the peptide-binding core (Lafuente & Reche, 2009). This suggests that the scoring function for the peptide that binds to the MHC complex is a linear sum of nine terms (Lund *et al.*, 2005). Therefore, a statistical model which would be able to handle the variability of the sequence length is needed.

Computer predictions of MHC-binding peptides can be used as a first screening method to determine what are the regions on the antigen that will induce a response and hidden Markov model (HMM) is an example of a stochastic model which has been used to solve this problem. Profile HMMs can be used to characterize similarities between protein sequences (Lund *et al.*, 2005). Homologous sequences may not share many identical amino acids and the question is how do we detect the similarities. Using information from multiple sequence alignment that reveals amino acid conservation, mutability and active sites, profile HMMs puts together all these information to produce the profile for the particular family of protein (Lund *et al.*, 2005). A question to be asked using HMM is whether a sequence belongs to a particular family or not. This study aims to study the application of profile HMM using dataset consisting of peptides that have a high binding affinity for MHC Class II complexes and to also describe the specificity of the model. The profile will also be represented using a sequence logo viewer (Schuster-Bockler *et al.*, 2004) which will provide a graphical representation of the states emitted by the profile HMM. The profile HMM is then investigated further by searching it against a sequence database. This is a method which could be used to eliminate peptides which do not fit the profile, thus reducing the number of potential vaccine candidates.

1.1.2 Objectives of the study

The objective of this project is to study and apply profile HMM-based approach of MHC-binding peptides from the MHCPEP database (Brusic *et al.*, 1998). The project will focus on familiarizing with the HMM-based approach using the HMMER3 software package (Finn *et al.*, 2011) and using dataset of peptides that significantly binds to two types of human MHC Class II molecule, HLA-DR1 and HLA-DR4.

1.1.3 Organization of this report

This report is organized as following; the introduction chapter is where I introduce the basic background of the problem being studied which is the prediction of MHC Class II binding sites or epitope binding sites as well as the method which I have chosen to tackle this problem, profile HMM-based approach. After that is the literature review chapter where I summarize some of the work that has been done in the prediction of MHC Class II binding sites, including using HMM-based approach. Next is the methods section where I explain in detail what are the steps I have taken, from pre-processing of dataset to multiple sequence alignment and building of profile HMM as well as the validation of the profile HMM. The subsequent chapter are the results where the results of the multiple sequence alignment, the profile HMM, the HMM logo or the sequence logo, as well as the results of the validation of the profile HMM is shown. This is then followed by the discussion of my findings and lastly is the conclusion.

Chapter 2

Literature Review

The immune system is divided into two types of responses; humoral and cell-mediated response. The humoral response targets not-self, free-floating objects, or antigens (foreign substances not recognized to be naturally in the body which will trigger an immune response) whilst cell-mediated response is triggered when a cell in the body is attacked by a pathogen which enters the cell, like a virus. The regulation of the immune response involves a molecule with the highest degree of polymorphism among mammalian proteins known as major histocompatibility complex (MHC) (Lund *et al.*, 2005).

The MHC complex regulates the process of immune response by recognizing the peptide that corresponds to the antigen (Mamitsuka, 1998) and are classified into three subgroups; MHC Class I, II and III. All the three subgroups of complexes are involved in certain pathways in the immune response where MHC Class I and II pathways mainly processes and presents antigens to T lymphocytes (Lund *et al.*, 2005). In the humoral response, when a lymphocyte (a type of leukocyte, or white blood cell) stumbles upon an antigen, it will surround and engulf the antigen. The cell's internal mechanisms will then chop up the antigen and present a part of the chopped up part on its surface. These cells are then called antigen presenting cells (APCs) and examples of APCs include macrophages, dendritic cells, and B-cells. For the purpose of this study, the focus will be on peptides that interact with the MHC Class II . Figure 2.1 explains a general response involving the MHC Class II complex when an antigen is detected.

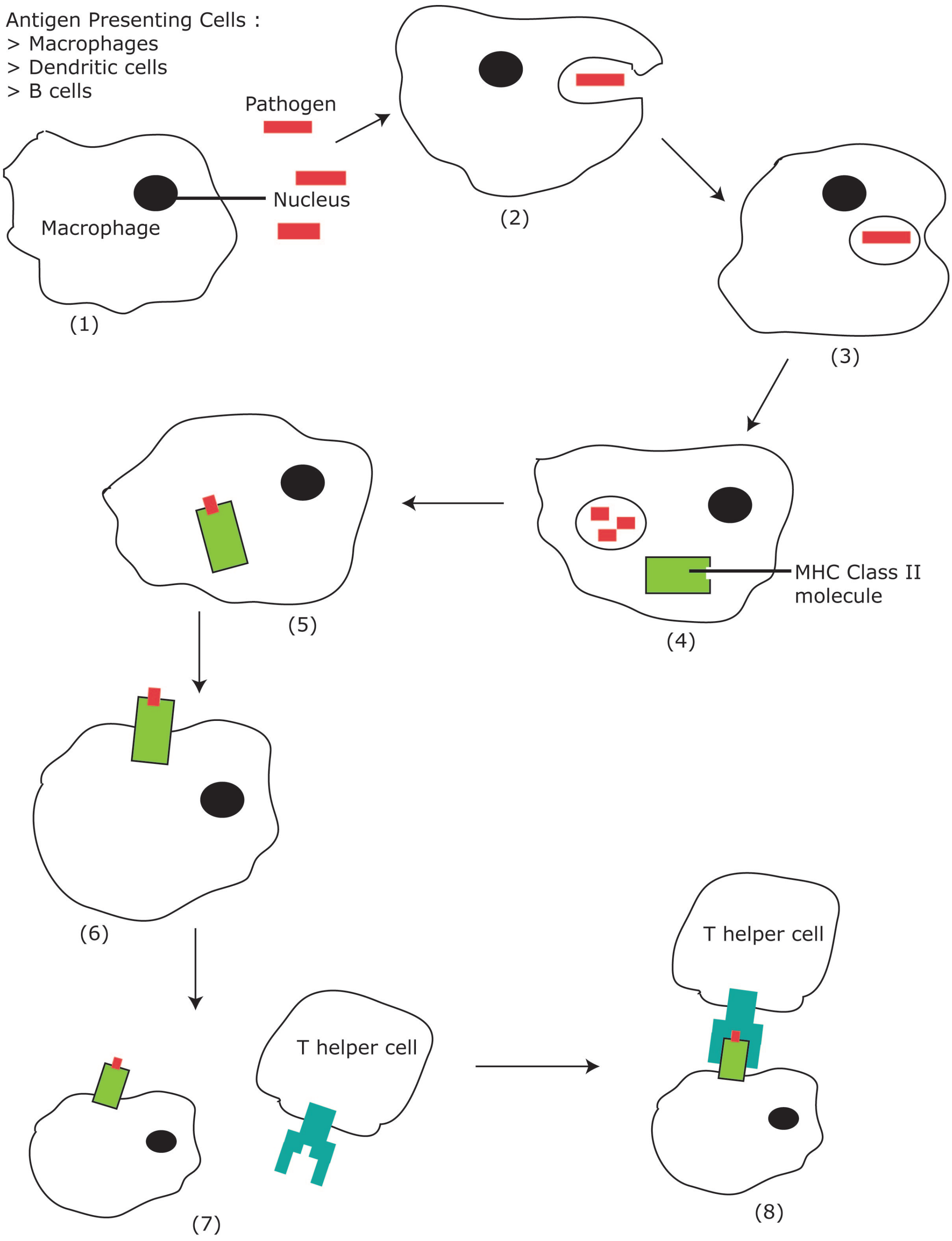


Figure 2.1: Diagram depicting the humoral response of the immune system. (1) A macrophage is an example of an antigen presenting cell. (2) Once it detects the pathogens, it will extend its membrane (3) to engulf the pathogen. (4) Once inside, reactions inside the macrophage will break the pathogen into fragments, disabling its function. Already present inside the macrophage is a MHC class II molecule. (5) The MHC class II molecule has a receptor that will recognize the complementary fragment from the pathogen and will form a MHC complex. (6) The complex will then fuse with the macrophage's membrane in order to 'present' the pathogenic fragment. (7) A T-helper cell with the complementary receptor to the MHC complex on the macrophage's membrane. (8) The T-helper cell will bind to the complex and this will induce the next chain of immunological response to neutralize the pathogens.

The human counterpart for the MHC molecules are called human leukocyte antigens (HLA). HLA Class II consists of several supertypes, and are categorized based on the shared binding characteristic. Examples of the supertypes under the HLA Class II is the HLA DR, DQ and DP. These supertypes also refer to the alleles which encode for the HLA molecules.

Investigating peptides that can bind to the MHC molecule, also known as epitope prediction, is an effective method which can be incorporated into the pipeline for vaccine design. Not all peptides will bind to the MHC molecule, and in fact, according to Mamitsuka (1998), it is said that only about 1 in 100-200 peptides will bind to the MHC molecule. Experimental methods like synthesizing the peptide and studying the binding activities can be a laborious and time-consuming process, hence computational methods can be used to facilitate the process of narrowing down potential candidates for vaccine design. One example of a website that provides tools for epitope prediction can be found at `http://www.immuneepitope.org/`.

Based on a review by Tong *et al.* (2006), several methods and protocols have been presented to predict immunogenic epitopes and they are generally classified into two main categories; methods that are based on the pattern identification in sequences of binding peptides and methods that model peptide with MHC interactions based on their three-dimensional (3D) structures. Examples of the former category would be hidden Markov models, artificial neural networks, procedures based on binding motifs, decision trees, and support vector machines. Examples of the latter category would include the use of homology modeling, docking and 3D techniques (Tong *et al.*, 2006).

Computer predictions of MHC Class II-binding peptides can be used as a first screening method to determine what peptide sequences will induce a response. One of the techniques which has been used in regards to prediction of MHC-binding peptides is hidden Markov model (HMM) (Mamitsuka, 1998; Noguchi *et al.*, 2002).

A Markov model consists of a set of states, in the case of biological sequence, those states can either be the four DNA nucleotides (A, C, G and T) or the twenty amino acids.

The outcome of an event in a Markov model depends only on the preceding state. Figure 2.2 uses DNA as an example for simplicity purposes.

Originally developed for speech recognition, HMMs have since been extensively used in computational gene finding. HMM is a stochastic model, and is suitable for the representation of time-series biological sequences with flexible lengths (Mamitsuka, 1998; Tong *et al.*, 2006). In biology, a hidden Markov model can be used to assign a state to each residue in a sequence (Lund *et al.*, 2005). An example is the B cell epitope model (Figure 2.3) where the positions in a protein is divided into two states; epitope or non-epitope. This information is kept hidden by the model where only the amino acid can be observed. The twenty amino acids have been identified to belong to three groups according to this model and they are hydrophobic [ACFILMPVW], uncharged polar [GNQSTY] and charged [DEHKR].

Profile HMM (pHMM) is a type of hidden Markov model that fits the modelling of multiple alignments (Durbin *et al.*, 1998) and is used to characterize sequence similarities within a family of proteins using the multiple sequence alignment of the protein sequences. One criteria of profile HMM is to take into account the gaps that occur at each of the position in the alignment. Profile HMMs also takes in the information of where these gaps are more or less likely to occur from the alignment (Durbin *et al.*, 1998). The probability of a given residue existing at a particular position in the alignment is assigned an emission probability, and the probability of the gaps provides position sensitive gap scores. The approach taken to building a profile HMM as described by Durbin *et al.* (1998) is the HMM is first built with repetitive states called the match states and all the transition probability is assigned as one. When dealing with gaps, insertions and deletions are treated individually. Sections of the sequence that do not match the model is introduced as an insert state and sections in the alignment that do not match any of the residues in a sequence is known as a delete state. Figure 2.4 shows a full profile HMM with the three states; match, insert and delete. Between these three states, transition can occur between two match states (m → m), a match and an insert (m → i), a match and a delete (m → d), an insert and a match (i → m).

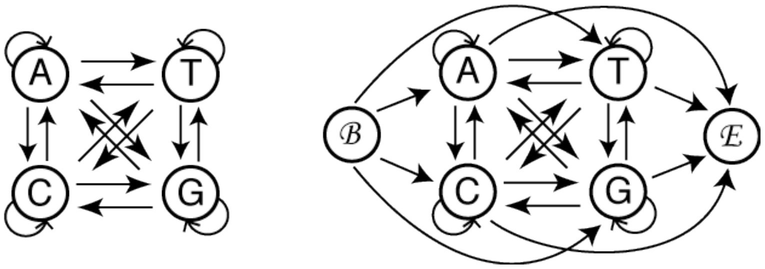


Figure 2.2: On the left is a basic Markov model using DNA as the states. Transition from one state to another (depicted by the arrow) is assigned a transition probability. Beginning (*B*) and end (*E*) states are treated as silent states and are added for modeling both ends of a sequence. These states do not emit any sequence.

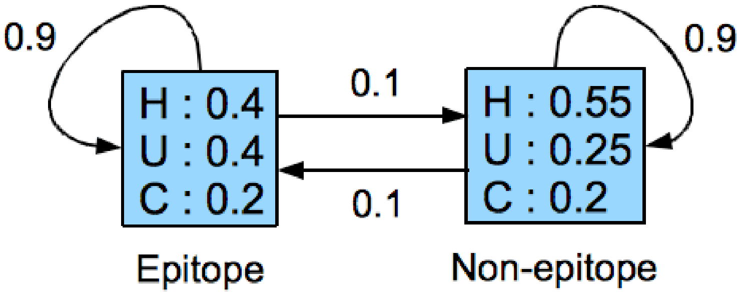


Figure 2.3: A B cell HMM model with two hidden states, epitope and non-epitope. The transition probability (represented by the arrow) is the probability of the epitope state transitioning to the non-epitope state or to itself, or the other way around. The observation is represented by the emission probability and this is given by the three groups of amino acid; hydrophobic (H), uncharged (U) and charged (C).

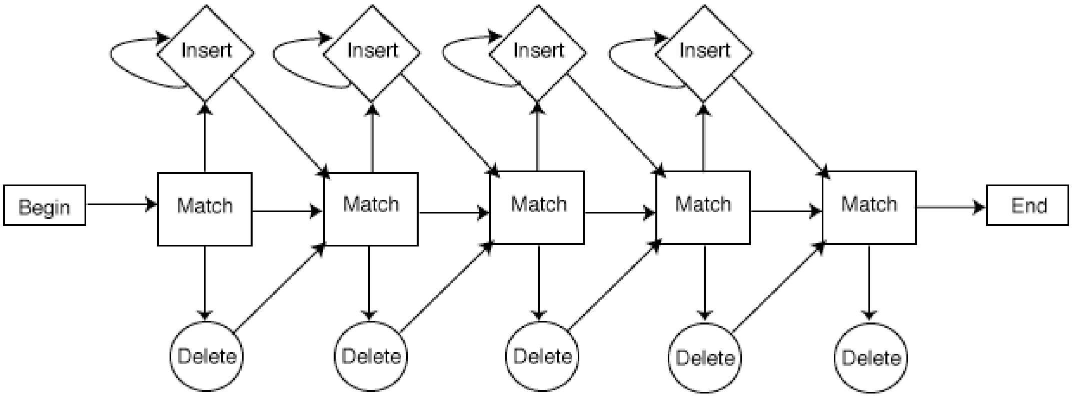


Figure 2.4: Profile hidden Markov model consisting of three states; match, insert and delete. The insert and match states have emission probabilities in general. Arrows from each state are assigned a transition probability. The begin and end state are silent states and only serves as a start and end point.

Probabilistic models such as this produces different outcomes with different probabilities and the parameters can be estimated from large sets of trusted examples called a training set (Durbin *et al.*, 1998). The training set can provide a reasonable estimate of the underlying probabilities of the model as long as it is not biased towards a particular residue composition. Profile HMMs can reveal important details such as amino acid conservation, mutations, and active sites (Lund *et al.*, 2005). Another application in biology using profile HMM suggests the removal of redundant sequences (with a 90% similarity over 90% of their length) from the training set (Singh *et al.*, 2009).

For this study, the software HMMER3, which is available for online use or download at `http://hmmer.janelia.org/` is used to train and build the HMM. HMMER3 takes in a multiple sequence alignment of DNA or proteins as input. It uses a heuristic algorithm called the multiple segment Viterbi (MSV) (Eddy, 2011) which is the dynamic programming algorithm doing the alignment of the sequence to the HMM.

Chapter 3

Method

3.1 Obtaining dataset and preprocessing of dataset

The MHC binding peptides dataset were downloaded from the Dana-Farber Repository for Machine Learning in Immunology (`http://bio.dfci.harvard.edu/DFRMLI/`). At the time of download, it contained 13,423 peptide sequences known to bind to MHC molecules. Information related to peptide sequence such as binding affinity (little, moderate, high) and MHC specificity (Class I, Class II) were also included in the dataset. For the purpose of this study, only peptides that were known to have a high binding affinity to MHC Class II were extracted using Perl. The peptides were then separated into its respective HLA group (Table 3.1). Two datasets were selected based on the higher number of sequences, HLA-DR1 (403 peptide sequences) and HLA-DR4 (313 peptide sequences). The peptide sequences were from different sources and some of them include organisms such as parasites (*Plasmodium falciparum*), phages, toxins (tetanus toxin) and viruses (hepatitis B virus).

Table 3.1: Dataset of peptides with positive binding to MHC Class II obtained from MHCPEP.

	Total sequence	High binding sequence
HLA-DR1	1761	403
HLA-DR2	326	34
HLA-DR4	1187	313
HLA-DR7	360	39
HLA-DR11	482	68
HLA-DR15	127	46
HLA-DR17	167	33
HLA-DR51	143	39
HLA-DQ2	115	47

Each dataset was then filtered to remove redundancy. Sequences more similar than 100%, 95% and 90% were eliminated (Table 3.2) using a web-based program called ElimDupes (http://hcv.lanl.gov/content/sequence/ELIMDUPES/elimdupes.html). For example, when a filtering parameter of 90% is set, if more than 90% of the length of a shorter sequence is covered by a larger sequence, it is considered a duplicate and removed from the dataset. There were no sequence set for sequences with more than 95% similarity for the HLA-DR1 dataset. This produces a total of seven sets of data to proceed with multiple sequence alignment, including the original sets of sequence with no filtering.

Table 3.2: Preprocessing of dataset

	Dataset filter (%)	n sequence
HLA-DR1	-	403
	100	309
	90	206
HLA-DR4	-	313
	100	251
	95	248
	90	175

3.2 Multiple sequence alignment and building of profile HMM

Multiple sequence alignment was carried out using ClustalX (Larkin *et al.*, 2007). The protein weight matrix used was the default matrix, the Gonnet series. Each of the alignment file was then used to build a profile HMM using hmmbuild in HMMER3. The alignment file was in fasta format and this had to be specified when running the hmmbuild program in the terminal:

```
$hmmbuild --informat afa <output.hmm> <input.fasta>
```

hmmbuild will produce an output file which can be viewed using a text editor. The .hmm file is divided into two regions; the first containing textual information and miscellaneous parameters and the second containing a tabular format for the main model parameters.

The consensus sequence was derived using the hmmemit program where the consensus parameter was stated in the command line at the terminal:

```
$hmmemit -c <hmmfile>
```

3.3 Evaluation of profile HMM using specificity study

For each profile HMM, a specificity study was carried out to determine whether the profile HMM could identify the sequences that would fit the model. This was done using the hmmsearch program in HMMER3. This program searches a profile against a sequence database. For the purpose of the specificity study, a dummy sequence database was generated for each profile HMM. Each dummy database contained 500 peptide sequences, 450 of which were considered to be synthetic true positives (TP) generated using the hmmemit program in HMMER3. This program samples sequences from the profile HMM and outputs them. The number of sequences were specified when running the hmmemit program in the terminal:

```
        $hmmemit -N 450
```

The remaining fifty sequence were considered as true negatives (TN) and were taken from the MHCPEP dataset. The peptides selected were peptide sequences recorded as peptides with little binding to MHC molecules. The profile HMM and the corresponding dummy sequence database were then used in hmmsearch:

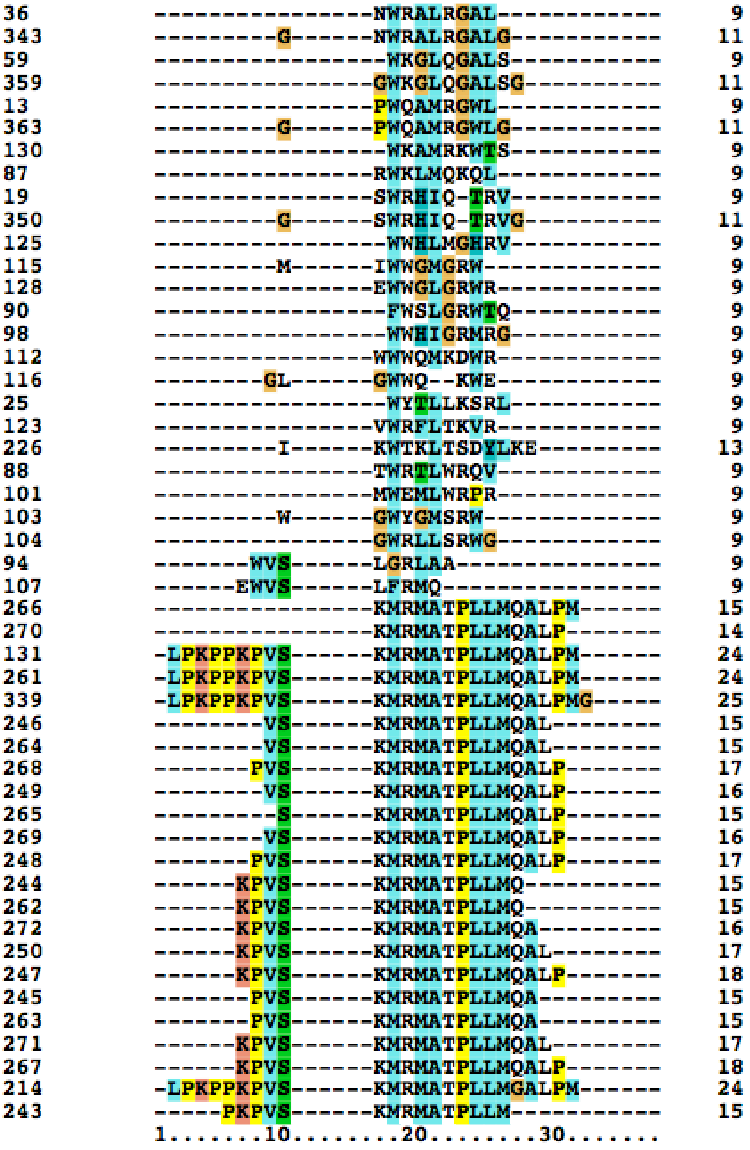
```
        $hmmsearch <hmmfile> <seqdb>
```

The output ranks the list of sequences with the most significant matches to the profile. The specificity of the profile HMM was then calculated using the ratio of the TN against the total of TN and FP (false positives) as a probability that the sequence has little binding affinity to the MHC molecule when it is a sequence with little binding affinity to the MHC molecule.

Chapter 4

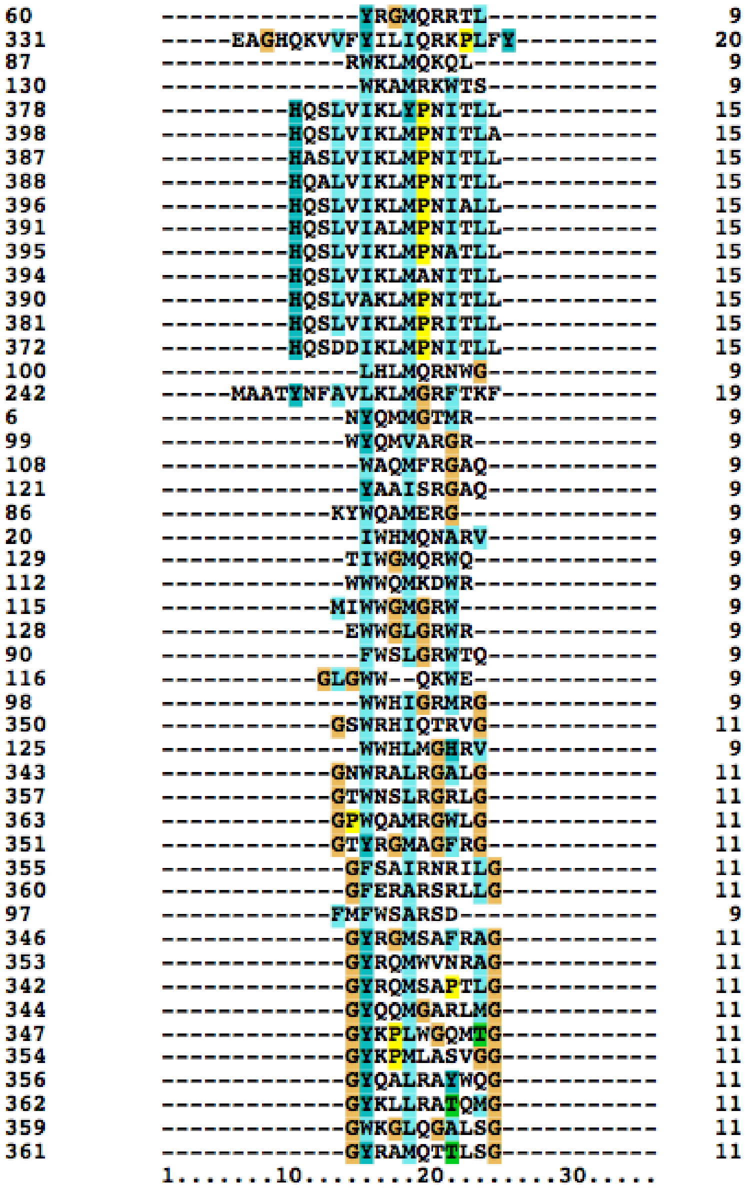
Results

Figure 4.1 is a section of the multiple sequence alignment for the DR1-od dataset (403 peptide sequences). This dataset came from the HLA-DR1 group and is the alignment of the sequences with no filtering done (the naming convention DR1-od is used to refer to the original dataset). Figure 4.2 is a section of the multiple sequence alignment for the DR1-90 dataset (206 peptide sequences). This dataset came from the HLA-DR1 group and is the alignment of the sequences with the filtering parameter of 90% (the naming convention DR1-90 is used for this set). This means that if more than 90% of the length of a shorter sequence is covered by a larger sequence, it is considered a duplicate and removed from the dataset before it is aligned. Figure 4.3 is a section of the multiple sequence alignment for the DR4-od dataset (313 peptide sequences). This dataset came from the HLA-DR4 group and is the alignment of the sequences with no filtering done. Figure 4.4 is a section of the multiple sequence alignment for the DR4-90 (175 peptide sequences) dataset. This dataset came from the HLA-DR4 group and is the alignment of the sequences with the filtering parameter of 90%. For full figures of the multple sequence alignments, refer to Supplementary Material.



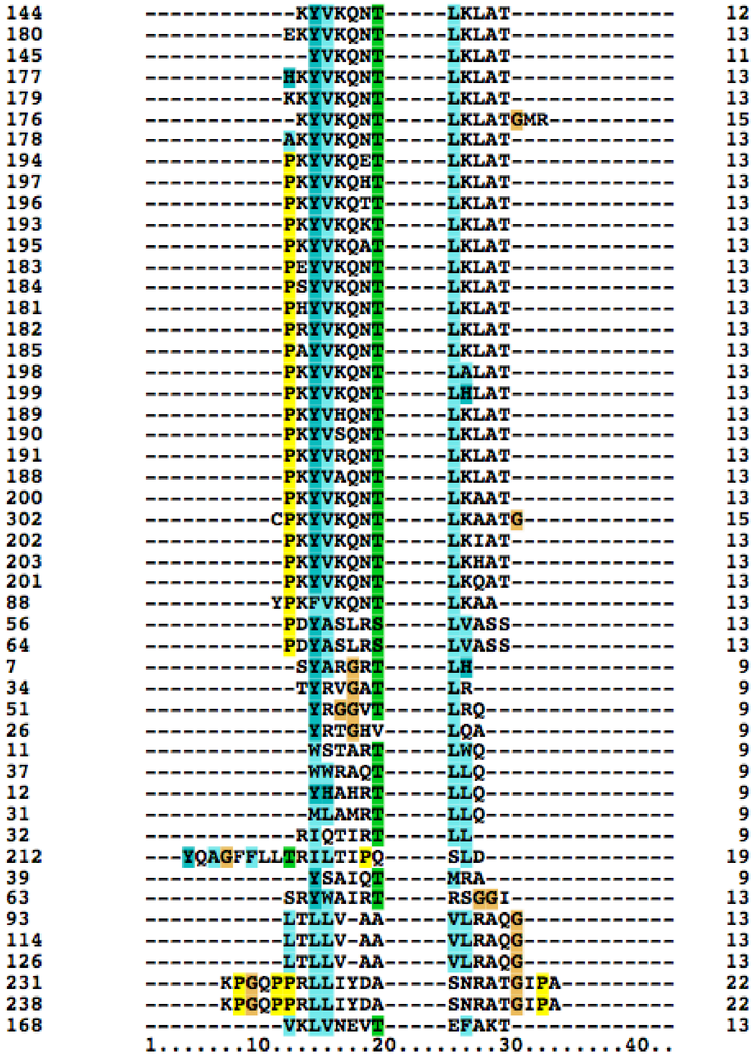
```
36          ----------------NWRALRGAL------------     9
343         ---------G------NWRALRGALG-----------    11
59          -----------------WKGLQGALS-----------     9
359         ----------------GWKGLQGALSG----------    11
13          ----------------PWQAMRGWL------------     9
363         ---------G------PWQAMRGWLG-----------    11
130         -----------------WKAMRKWTS-----------     9
87          ----------------RWKLMQKQL------------     9
19          ----------------SWRHIQ-TRV-----------     9
350         ---------G------SWRHIQ-TRVG----------    11
125         -----------------WWHLMGHRV-----------     9
115         ---------M------IWWGMGRW-------------     9
128         ----------------EWWGLGRWR------------     9
90          -----------------FWSLGRWTQ-----------     9
98          -----------------WWHIGRMRG-----------     9
112         ----------------WWWQMKDWR------------     9
116         --------GL------GWWQ--KWE------------     9
25          -----------------WYTLLKSRL-----------     9
123         ----------------VWRFLTKVR------------     9
226         ---------I------KWTKLTSDYLKE---------    13
88          ----------------TWRTLWRQV------------     9
101         ----------------MWEMLWRPR------------     9
103         ---------W------GWYGMSRW-------------     9
104         ----------------GWRLLSRWG------------     9
94          -------WVS------LGRLAA---------------     9
107         ------EWVS------LFRMQ----------------     9
266         ----------------KMRMATPLLMQALPM------    15
270         ----------------KMRMATPLLMQALP-------    14
131         -LPKPPKPVS------KMRMATPLLMQALPM------    24
261         -LPKPPKPVS------KMRMATPLLMQALPM------    24
339         -LPKPPKPVS------KMRMATPLLMQALPMG-----    25
246         --------VS------KMRMATPLLMQAL--------    15
264         --------VS------KMRMATPLLMQAL--------    15
268         -------PVS------KMRMATPLLMQALP-------    17
249         --------VS------KMRMATPLLMQALP-------    16
265         ---------S------KMRMATPLLMQALP-------    15
269         --------VS------KMRMATPLLMQALP-------    16
248         -------PVS------KMRMATPLLMQALP-------    17
244         ------KPVS------KMRMATPLLMQ----------    15
262         ------KPVS------KMRMATPLLMQ----------    15
272         ------KPVS------KMRMATPLLMQA---------    16
250         ------KPVS------KMRMATPLLMQAL--------    17
247         ------KPVS------KMRMATPLLMQALP-------    18
245         -------PVS------KMRMATPLLMQA---------    15
263         -------PVS------KMRMATPLLMQA---------    15
271         ------KPVS------KMRMATPLLMQAL--------    17
267         ------KPVS------KMRMATPLLMQALP-------    18
214         -LPKPPKPVS------KMRMATPLLMGALPM------    24
243         -----PKPVS------KMRMATPLLM-----------    15
            1.......10........20........30.......
```

Figure 4.1: A section of the multiple sequence alignment for DR1-od



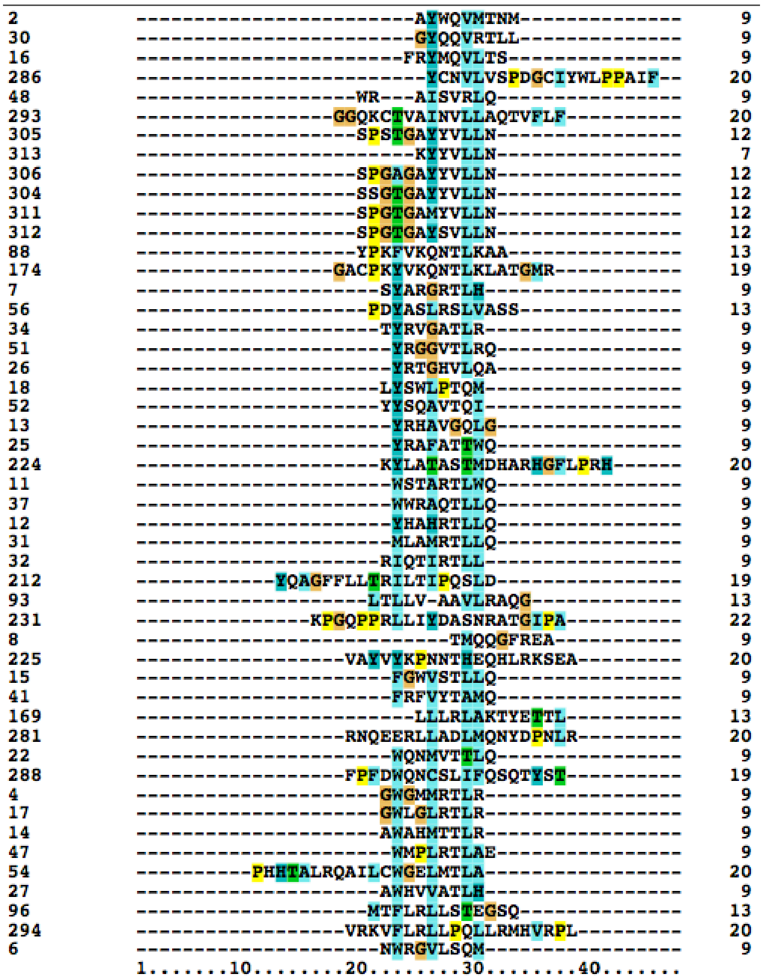
```
60          --------------YRGMQRRTL------------      9
331         -----EAGHQKVVFYILIQRKPLFY----------     20
87          -------------RWKLMQKQL-------------      9
130         --------------WKAMRKWTS------------      9
378         ---------HQSLVIKLYPNITLL-----------     15
398         ---------HQSLVIKLMPNITLA-----------     15
387         ---------HASLVIKLMPNITLL-----------     15
388         ---------HQALVIKLMPNITLL-----------     15
396         ---------HQSLVIKLMPNIALL-----------     15
391         ---------HQSLVIALMPNITLL-----------     15
395         ---------HQSLVIKLMPNATLL-----------     15
394         ---------HQSLVIKLMANITLL-----------     15
390         ---------HQSLVAKLMPNITLL-----------     15
381         ---------HQSLVIKLMPRITLL-----------     15
372         ---------HQSDDIKLMPNITLL-----------     15
100         --------------LHLMQRNWG------------      9
242         -----MAATYNFAVLKLMGRFTKF-----------     19
6           -------------NYQMMGTMR-------------      9
99          -------------WYQMVARGR-------------      9
108         --------------WAQMFRGAQ------------      9
121         --------------YAAISRGAQ------------      9
86          ------------KYWQAMERG--------------      9
20          --------------IWHMQNARV------------      9
129         -------------TIWGMQRWQ-------------      9
112         -------------WWWQMKDWR-------------      9
115         ------------MIWWGMGRW--------------      9
128         -------------EWWGLGRWR-------------      9
90          --------------FWSLGRWTQ------------      9
116         -----------GLGWW--QKWE-------------      9
98          --------------WWHIGRMRG------------      9
350         ------------GSWRHIQTRVG------------     11
125         --------------WWHLMGHRV------------      9
343         ------------GNWRALRGALG------------     11
357         ------------GTWNSLRGRLG------------     11
363         ------------GPWQAMRGWLG------------     11
351         ------------GTYRGMAGFRG------------     11
355         -------------GFSAIRNRILG-----------     11
360         -------------GFERARSRLLG-----------     11
97          ------------FMFWSARSD--------------      9
346         -------------GYRGMSAFRAG-----------     11
353         -------------GYRQMWVNRAG-----------     11
342         -------------GYRQMSAPTLG-----------     11
344         -------------GYQQMGARLMG-----------     11
347         -------------GYKPLWGQMTG-----------     11
354         -------------GYKPMLASVGG-----------     11
356         -------------GYQALRAYWQG-----------     11
362         -------------GYKLLRATQMG-----------     11
359         -------------GWKGLQGALSG-----------     11
361         -------------GYRAMQTTLSG-----------     11
            1.......10........20........30.....
```

Figure 4.2: A section of the multiple sequence alignment for DR1-90



```
144   -------------KYVKQNT-----LKLAT------------   12
180   ------------EKYVKQNT-----LKLAT------------   13
145   --------------YVKQNT-----LKLAT------------   11
177   ------------HKYVKQNT-----LKLAT------------   13
179   ------------KKYVKQNT-----LKLAT------------   13
176   -------------KYVKQNT-----LKLATGMR---------   15
178   ------------AKYVKQNT-----LKLAT------------   13
194   ------------PKYVKQET-----LKLAT------------   13
197   ------------PKYVKQHT-----LKLAT------------   13
196   ------------PKYVKQTT-----LKLAT------------   13
193   ------------PKYVKQKT-----LKLAT------------   13
195   ------------PKYVKQAT-----LKLAT------------   13
183   ------------PEYVKQNT-----LKLAT------------   13
184   ------------PSYVKQNT-----LKLAT------------   13
181   ------------PHYVKQNT-----LKLAT------------   13
182   ------------PRYVKQNT-----LKLAT------------   13
185   ------------PAYVKQNT-----LKLAT------------   13
198   ------------PKYVKQNT-----LALAT------------   13
199   ------------PKYVKQNT-----LHLAT------------   13
189   ------------PKYVHQNT-----LKLAT------------   13
190   ------------PKYVSQNT-----LKLAT------------   13
191   ------------PKYVRQNT-----LKLAT------------   13
188   ------------PKYVAQNT-----LKLAT------------   13
200   ------------PKYVKQNT-----LKAAT------------   13
302   -----------CPKYVKQNT-----LKAATG-----------   15
202   ------------PKYVKQNT-----LKIAT------------   13
203   ------------PKYVKQNT-----LKHAT------------   13
201   ------------PKYVKQNT-----LKQAT------------   13
88    -----------YPKFVKQNT-----LKAA-------------   13
56    ------------PDYASLRS-----LVASS------------   13
64    ------------PDYASLRS-----LVASS------------   13
7     -------------SYARGRT-----LH---------------    9
34    -------------TYRVGAT-----LR---------------    9
51    --------------YRGGVT-----LRQ--------------    9
26    --------------YRTGHV-----LQA--------------    9
11    --------------WSTART-----LWQ--------------    9
37    --------------WWRAQT-----LLQ--------------    9
12    --------------YHAHRT-----LLQ--------------    9
31    --------------MLAMRT-----LLQ--------------    9
32    -------------RIQTIRT-----LL---------------    9
212   ---YQAGFFLLTRILTIPQ------SLD--------------   19
39    --------------YSAIQT-----MRA--------------    9
63    ------------SRYWAIRT-----RSGGI------------   13
93    ------------LTLLV-AA-----VLRAQG-----------   13
114   ------------LTLLV-AA-----VLRAQG-----------   13
126   ------------LTLLV-AA-----VLRAQG-----------   13
231   ------KPGQPPRLLIYDA------SNRATGIPA--------   22
238   ------KPGQPPRLLIYDA------SNRATGIPA--------   22
168   ------------VKLVNEVT-----EFAKT------------   13
      1.......10........20........30........40..
```

Figure 4.3: A section of the multiple sequence alignment for DR4-od



```
2       ------------------------AYWQVMTNM--------------   9
30      ------------------------GYQQVRTLL--------------   9
16      -----------------------FRYMQVLTS---------------   9
286     -------------------------YCNVLVSPDGCIYWLPPAIF--  20
48      -------------------WR---AISVRLQ----------------   9
293     -----------------GGQKCTVAINVLLAQTVFLF----------  20
305     -------------------SPSTGAYYVLLN----------------  12
313     ------------------------KYYVLLN----------------   7
306     -------------------SPGAGAYYVLLN----------------  12
304     -------------------SSGTGAYYVLLN----------------  12
311     -------------------SPGTGAMYVLLN----------------  12
312     -------------------SPGTGAYSVLLN----------------  12
88      -------------------YPKFVKQNTLKAA---------------  13
174     -----------------GACPKYVKQNTLKLATGMR-----------  19
7       ---------------------SYARGRTLH-----------------   9
56      --------------------PDYASLRSLVASS--------------  13
34      ---------------------TYRVGATLR-----------------   9
51      ----------------------YRGGVTLRQ----------------   9
26      ----------------------YRTGHVLQA----------------   9
18      ---------------------LYSWLPTQM-----------------   9
52      ---------------------YYSQAVTQI-----------------   9
13      ----------------------YRHAVGQLG----------------   9
25      ----------------------YRAFATTWQ----------------   9
224     ---------------------KYLATASTMDHARHGFLPRH------  20
11      ----------------------WSTARTLWQ----------------   9
37      ----------------------WWRAQTLLQ----------------   9
12      ----------------------YHAHRTLLQ----------------   9
31      ----------------------MLAMRTLLQ----------------   9
32      ---------------------RIQTIRTLL-----------------   9
212     ------------YQAGFFLLTRILTIPQSLD----------------  19
93      --------------------LTLLV-AAVLRAQG-------------  13
231     ---------------KPGQPPRLLIYDASNRATGIPA----------  22
8       ---------------------------TMQQGFREA-----------   9
225     ------------------VAYVYKPNNTHEQHLRKSEA---------  20
15      ----------------------FGWVSTLLQ----------------   9
41      ----------------------FRFVYTAMQ----------------   9
169     ------------------------LLLRLAKTYETTL----------  13
281     ------------------RNQEERLLADLMQNYDPNLR---------  20
22      ----------------------WQNMVTTLQ----------------   9
288     ------------------FPFDWQNCSLIFQSQTYST----------  19
4       ---------------------GWGMMRTLR-----------------   9
17      ---------------------GWLGLRTLR-----------------   9
14      ---------------------AWAHMTTLR-----------------   9
47      ----------------------WMPLRTLAE----------------   9
54      ----------PHHTALRQAILCWGELMTLA-----------------  20
27      ---------------------AWHVVATLH-----------------   9
96      --------------------MTFLRLLSTEGSQ--------------  13
294     ------------------VRKVFLRLLPQLLRMHVRPL---------  20
6       ---------------------NWRGVLSQM-----------------   9
        1.......10........20........30........40.......
```

Figure 4.4: A section of the multiple sequence alignment for DR4-90

Table 4.1 is a summary of the profile HMM for all seven datasets. The profile HMM was named according to the HLA group it belongs to as well as the filtering parameter used for that dataset. The aligned column was derived from the multiple sequence alignment which HMMER turned into a model of consensus positions. The difference between aligned column and consensus sequence would be the gap-containing alignment columns to be insertions relative to consensus.

Table 4.2 shows the consensus sequence derived from each of the profile HMM. The residue distribution for hydrophobic (ACFILMPVW), uncharged polar (GNQSTY) and charged (DEHKR) residues were calculated for each of the consensus sequence.

Figure 4.5 - figure 4.8 is the profile HMM and sequence logo for each of the dataset. The profile HMM was viewed using HMMVE v1.2 : A Visual Editor for Profile Hidden Markov Model (Dai & Cheng, 2008). Below the profile HMM is the sequence logo using the online version of LogoMAT-M (`http://www.sanger.ac.uk/cgi-bin/software/analysis/logomat-m.cgi`). Sequence logos are used to illustrate the content and distribution of a multiple alignment. LogoMAT-M is a type of sequence logo used to view the distribution of a profile HMM. In the case of a protein alignment, the residues are shown as a stack of letters. The letter at the top of the stack has the highest probability, and the probability reduces going down the stack. The size of the letter also contains information; it depicts the emission probability from the distribution of the state. The width of the column shows the relative contribution of the position to the overall model. The pink columns will most probably not contain any letters and these are the insert states. The residues in LogoMAT-M are color-coded according to their biological properties; charged [DEHKR], polar, uncharged [CNPQST], aliphatic [AGILVM] and aromatic [FWY].

Table 4.1: Summary of profile HMM for all seven datasets. The aligned column was derived from the multiple sequence alignment which HMMER turned into a model of consensus positions.

Profile HMM	n sequence	Dataset filter (%)	Aligned column	Consensus sequence	Gap-containing alignment
DR1-od	403	–	37	21	16
DR1-100	309	100	34	24	10
DR1-90	206	90	35	22	13
DR4-od	313	–	42	34	8
DR4-100	251	100	44	38	6
DR4-95	248	95	45	36	9
DR4-90	175	90	47	41	6

Table 4.2: Consensus sequence derived from the profile HMMs. The residue distribution column describes the content of hydrophobic [ACFILMPVW], uncharged polar [GNQSTY] and charged [DEHKR] residues in the sequence.

Profile HMM	Consensus sequence	Sequence length	Residue distribution		
			Hydrophobic	Uncharged polar	Charged
DR1-od	KPVSPKYRLALPLLMQALPMK	21	14	3	4
DR1-100	AAPPLDILVLTGMKARLLPILPPM	24	19	2	3
DR1-90	ALAPDVLFFRKMLALRLLPQLL	22	17	1	4
DR4-od	GIVEQSSLSISLFVNANLLNSALFYCPIAIMAAL	34	22	11	1
DR4-100	PHHAAARQAQPLYKASNNVLLANANIFYCPIAIMSALA	38	24	10	4
DR4-95	PHHTAARQALPAFLAALAAALASPDGCIYWLPPAIF	36	27	5	4
DR4-90	ENGEWAIQHRPAKMPALDALSAAAAQKALGLTGSFLAKGPS	41	22	11	8

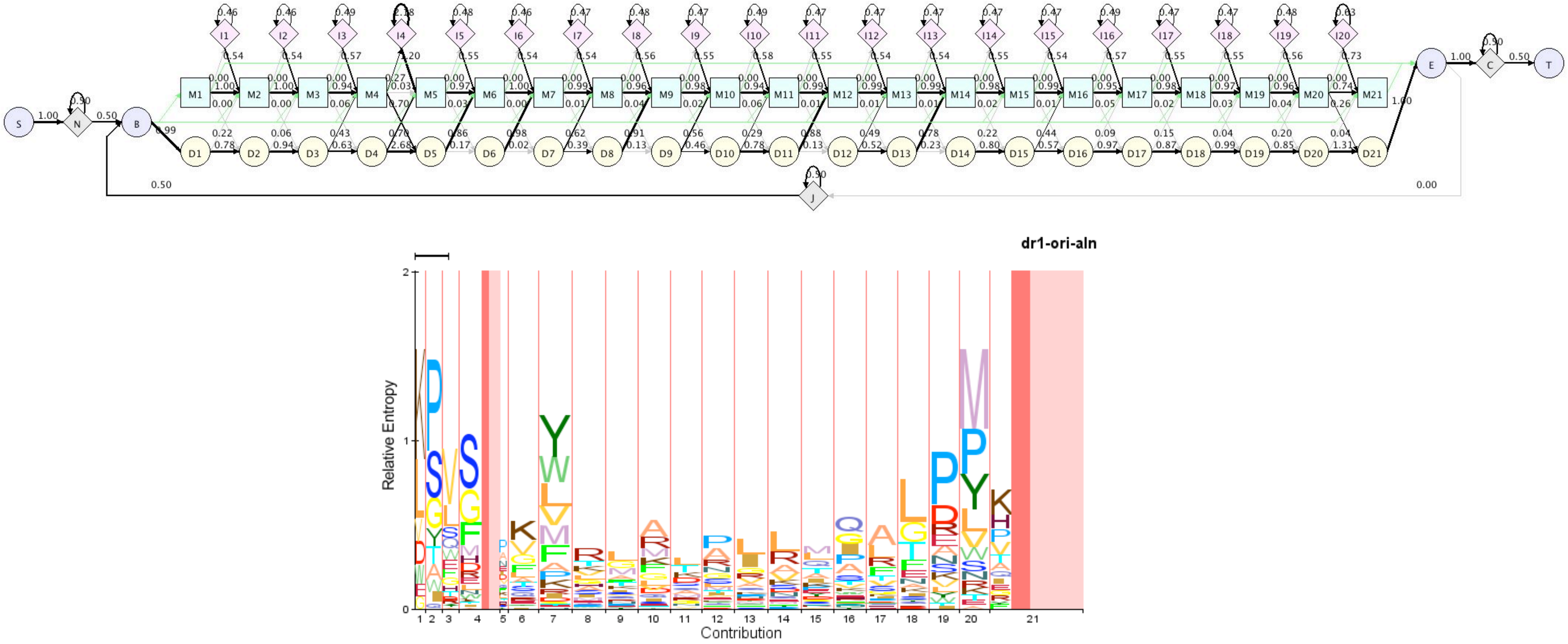


Figure 4.5: Profile HMM and sequence logo of DR1-od

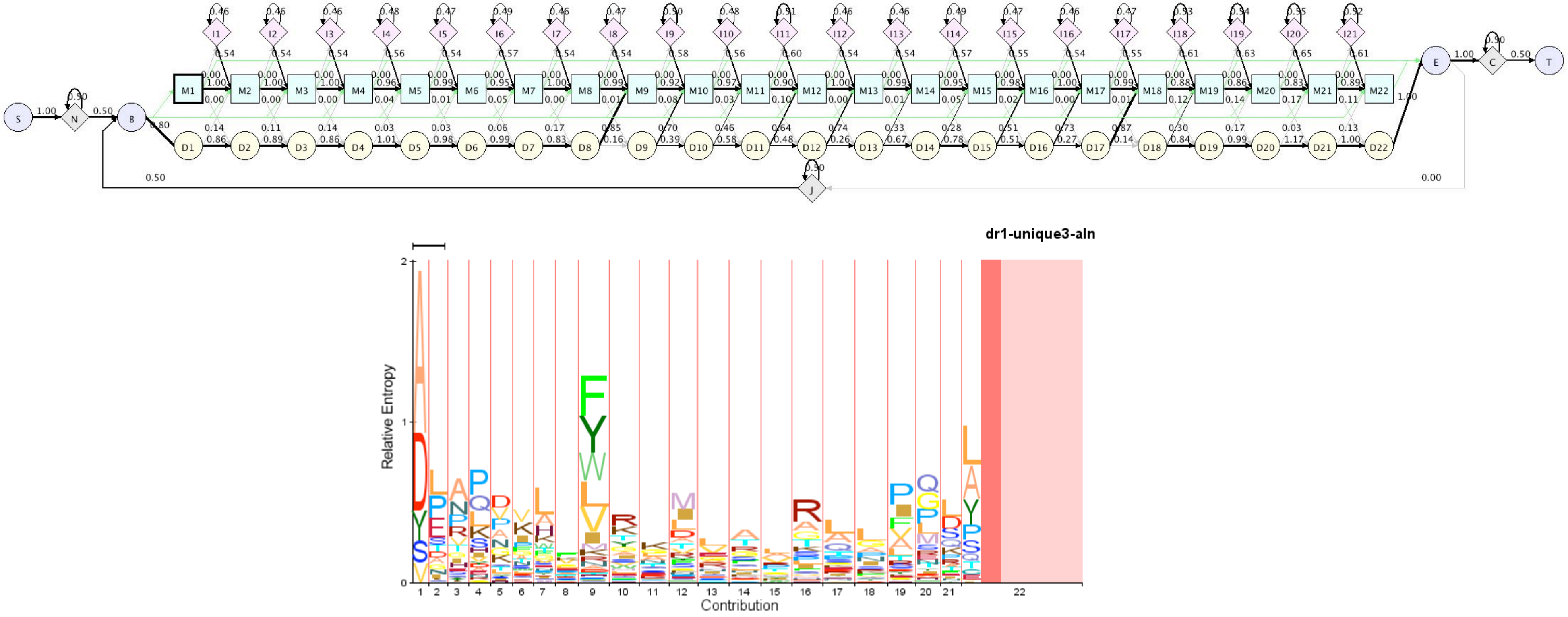


Figure 4.6: Profile HMM and sequence logo of DR1-90

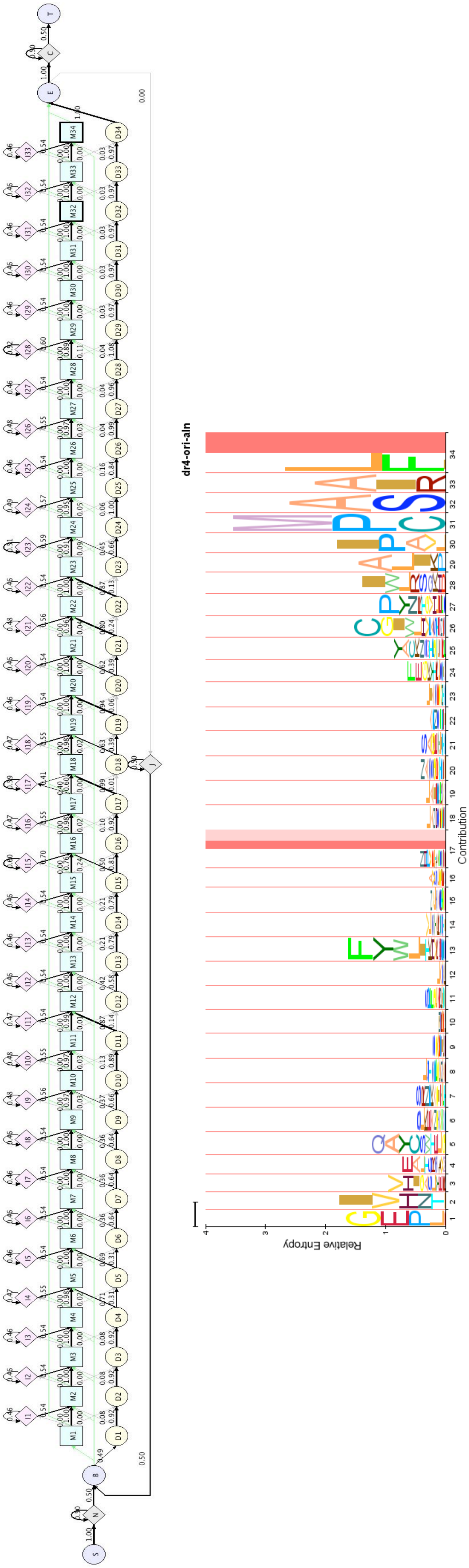


Figure 4.7: Profile HMM and sequence logo of DR4-od

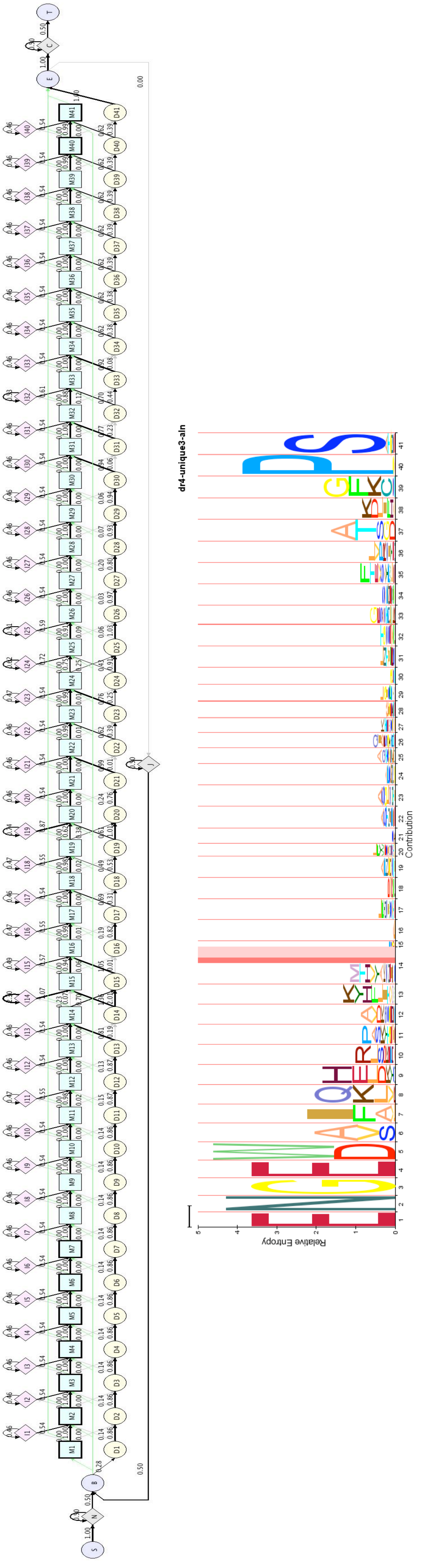


Figure 4.8: Profile HMM and sequence logo of DR4-90

Table 4.2 is the result of using hmmsearch to evaluate each profile HMM. The number of targets above threshold is the number of sequences from the sequence database (the dummy dataset consisting of 450 synthetic true positives and 50 true negatives) that the profile HMM was searched against that made it through HMMER3's filters. Based on the two E-values produced in the output of hmmsearch, if both conditional and independent E-values were significant (less than 1), the sequence is likely to be homologous to the query. All the targets from the dummy dataset which were above the threshold had E-values which were significant. From here, the specificity study was carried out, and each profile HMM scored a probability of one.

Table 4.3: Evaluation of profile HMM using hmmsearch

Dataset	Number of targets above threshold
DR1-od	11
DR1-100	4
DR1-90	3
DR4-od	297
DR4-100	361
DR4-95	398
DR4-90	411

Chapter 5

Discussion

The profile HMM method relies on the quality of the training data being used (Durbin *et al.*, 1998) as well as the quality of the multiple sequence alignment (Eddy, 1998). The scoring method is based on the true frequency of a residue at a given position in the alignment from its observed frequency. Therefore, redundancies within the dataset must be eliminated in order to avoid bias towards the composition of a particular residue at a particular position.

The HLA-DR1 group was divided according to two different filters; 100% (DR1-100) and 90% (DR1-90). The HLA-DR4 group was divided according to three different filters; 100% (DR4-100), 95% (DR1-95) and 90% (DR4-90). Each dataset, including the original dataset (DR1-od and DR4-od), was submitted for a multiple sequence alignment and from there the profile HMM was built. Despite coming from the same group, DR1-od, DR-100 and DR1-90 produced different profile HMMs. The number of consensus sequence vary slightly (21 for DR1-od, 24 for DR1-100, 22 for DR1-90). These consensus positions also correlate with the match states of the profile HMM. This shows that the training set as well as the multiple sequence alignment will affect the building of the profile HMM due to the contribution of a particular residue at a particular position. Same goes for the HLA-DR4 group. The number of consensus sequence vary slightly as well (34 for DR4-od, 38 for DR4-100, 36 for DR4-95, 41 for DR4-90). Another trend which can be observed is the number of gap-containing alignment. The gap-containing alignment decreases when more stringent parameters are used. For the HLA-DR1 group, the DR1-od has 16 gaps, DR1-100 has 10 gaps, DR1-90 has 13 gaps. For the HLA-DR4 group, the DR4-od has 8 gaps, DR1-100 has 6 gaps, DR4-95 has 9 gaps, DR-90 has 6 gaps.

Based on the consensus sequence of all the datasets, the residue distribution for hydrophobic, uncharged polar and polar was calculated. All seven consensus sequences showed a high content of hydrophobic residues. For datasets in each of the HLA group (HLA-DR1 and HLA-DR4), although the consensus sequence differed from one another, the content for hydrophobic residues is the highest for all the datasets.

Based on the specificity study carried out, it was apparent that the profile HMM is able to differentiate between the true positive sequences with the true negatives. In fact, the specificity study carried out [TN $\div$ (TN + FP)] showed that all seven datasets had a probability of one. From the sequence targets identified to be homologous to the profile, none of the true negative was in the output. A possible explanation for this would be due to the fact that the true positives were generated based on the profile HMM model itself. These results might not reflect a realistic situation due to the bias. Possible future work to rectify this would be to increase the number of dataset to include actual validated experimental data so as to improve the reliability of the test dataset.

Due to the fact that the profile HMM method is still a statistical method, it is important to determine the biology knowledge behind the application. A profile HMM can always be built using a multiple sequence alignment, but more information is needed to make it a meaningful one. Letting the model train itself can be used at a preliminary stage of analysis, but must later on be incorporated with the biological knowledge relevant to the problem being asked. Information such as structural characteristic of the amino acid can be included to build a better profile HMM. The MHC Class II binding groove has been proposed to consist of only 9 residues. In order to provide a better profile HMM for MHC Class II binding peptides, these information should be incorporated so as to produce more realistic results.

Chapter 6

Conclusion

Based on the basic analysis done on HMM profiles generated by HMMER3 on the two group of MHC-binding dataset (HLA-DR1 and HLA-DR4), there is still much to be investigated. Most importantly is the biological background of the profile HMM for MHC-binding peptides. Further investigation of the correlation between biological properties of the amino acid such as hydrophobic, aliphatic and uncharged polar of the peptides need to be studied. An example is the B cell HMM model where the basic distribution of the amino acids have been characterized (Figure 2.3).

Profile HMM is a good method for finding similarities between distantly related sequences of varying lengths. An important thing to note is the training data used for building the profile HMM; the training set can provide a reasonable estimate of the underlying probabilities of the model provided that it is not biased towards a particular residue composition.

In order for this method to be included in the pipeline for vaccine design, a more careful analysis needs to be done. Based on this preliminary study, it has the potential to be investigated further for future research.