Prediction of MHC Class II binding peptide, Using Support Vector Machines

Jia Mi [*]

Department of Biochemistry and Biophysics, Stockholm University

Supervisor: Arne Elofsson, Stockholm Bioinformatics Center,Stockholm University

April 16, 2003

[*] Stockholm Bioinformatics Center, Stockholm University, SE-106 91 Stockholm, Sweden E–mail: jia@sbc.su.se

Abstract

Major histocompatibility complex (MHC) molecules play a critical role in initiating and regulation immune responses. Helper T lymphocytes can recognize a complex formed between a MHC class II molecule and an antigenic peptide. Determining which peptides bind to a specific MHC molecule is fundamental to understanding the basis of immunity, and for the development of vaccines and immunotherapeutics for autoimmune diseases and cancer.

This Master's thesis investigates the usage of Support Vector Machine for MHC class II binding peptide prediction. For each allele a *model cluster* (SVMHCII) which contains 20 models was created. The training data wass obtained from the public database MHCPEP. For the 26 different alleles that contain enough data we obtained average Mc coefficients of 0.57-0.74. A comparison between SVMHCII and a public MHC II predictor ProPred is made. For 10 out of 11 alleles SVMHCII perform better than ProPred. Generally, these results indicate that SVMHCII has a strong ability for MHC class II binding prediction.

Contents

1 Introduction **4**

2 Background and Theory **8**
2.1 Immune system . . . 8
2.1.1 MHC Cells involved in acquired immunity and specific recognition . . . 8
2.2 Pattern recognition and machine learning . . . 9
2.2.1 Separable patterns . . . 9
2.2.2 Machine Learning . . . 11
2.3 Support Vector Machine . . . 12
2.4 Cross-validation . . . 13
2.5 Performance measurements . . . 13

3 Material and Methods **15**
3.1 MHC peptide databases . . . 15
3.1.1 MHCPEP: MHC binding database . . . 15
3.1.2 ENSEMBLE: non-binding database . . . 15
3.2 SVM^{light} . . . 15
3.2.1 SVM_learn and SVM_classify . . . 16
3.2.2 Parameters in SVM$^{\mathrm{light}}$ . . . 16
3.3 Algorithm for prediction . . . 16
3.4 Preprocessing the database . . . 17
3.4.1 Initialization . . . 18
3.4.2 Model establishment and cluster . . . 19
3.4.3 The prediction using *model clusters* . . . 19
3.5 Comparison with ProPred . . . 20
3.6 Test-set validation . . . 20

4 Results and Discussion **22**
4.1 Results from Model clusters . . . 22
4.2 Comparison with ProPred . . . 22
4.3 Test-set validation . . . 23

5 Conclusions and future steps **24**
5.1 Conclusions . . . 24
5.2 Future development . . . 26

6 Acknowledgments **26**

7 References **27**

1 Introduction

The process of binding between Major Histocompatibility Complex (MHC) and the peptides plays an important role in the immune system. To predict the binding ability of peptides to a specific MHC molecule is fundamental to understanding the basis of immunity, and for the development of vaccines and amino-therapeutics for autoimmune disease and cancer. The immune system is a defense system that is present in vertebrates to protect them from invading pathogens. One of the most important parts of the immune system is the specific recognition of antigens bound to major Histocompatibility molecules carried out by T-cells.

There is a natural turnover of proteins in living cells, which means that they are hydrolyzed into smaller peptide fragments. Some of these peptides bind to MHC molecules and travel to the cell surface, where the MHC-peptide complex can be recognized by T-cell receptors (TCRs) on T-cells. The presentation of MHC-peptide complexes is a way to monitor what is going on in the body. If there is no foreign antigen in the body, MHC presents only self-peptides and hence there is no activation of T-cells. On the other hand, if the peptide presented by MHC is, for example, a viral protein, T-cells can be activated.

Cytotoxic T-cells recognize peptides bound to MHC molecules. These MHC-peptide complexes are potential tools for the diagnosis and control of pathogens and cancer. One major problem is to find out what peptides from a protein that actually bind to a MHC molecule. Suggestions have been made that only 1 in 100-200 possible binders actually do bind. A reliable prediction method that reduces the number of candidate binders is therefore useful. This paper presents a method for predicting peptides that bind MHC class II using Support Vector Machines. Sequence data of peptides that bind different MHC types were extracted from the public database MHCPEP.

The prediction of MHC II binding peptides is much more difficult than that of MHC I because of the difference of the structures of MHC I and MHC II. Figure 1 shows the main difference of MHC I and II peptides binding. Among the difficulties that must be addressed are: (i) the variable lengths of reported binding peptides; (ii) the undetermined core regions for individual peptides; (iii) the number of amino acids permissible as primary anchors; (iv) the range of experimental methods for assaying of peptide binding; (v) the experimental and reporting errors.

Prediction of peptides that bind to MHC can be divided into two groups: sequence based and structure based. Peptide binding to MHC is allele specific. By looking at frequencies of different amino acids in different positions for a large number of known binders, sequence motifs can be seen.

Figure 1: The mainly difference in structure between MHC I and MHC II. For MHC I (left)the both ends of binding groove are closed while for MHC II (right)they are open.

An example of a sequence motif might be the one seen for peptides that bind to a MHC I molecule HLA-A0201. It is very common that peptides that bind to this MHC molecule have a lysine in position 2 and a valine in position 9, the length is often 9 amino acids. Sequence motifs like this can be used as a simple prediction method [1]. More information can also be added to create a scoring matrix, in which every column correspond to a certain position in the peptide and the rows corresponds to the amino acids. Another approach for prediction is based on structural information about MHC-peptide complexes and evaluates how well a new peptide fits in the binding groove of a MHC molecule. A new peptide is threaded through a structural template to obtain a rough estimate of the binding energy. The energy estimation is based on the interactions seen in a solved crystal structure.

Prediction has also been made by using machine learning approaches such as artificial neural networks and hidden Markov models. The main feature of machine learning in this case is that they seem to reduce the number of positive false results (FP) compared to motif based methods. The importance of secondary anchors and deleterious residues at non-conserved regions places limitations on the usefulness of motifs.

The prediction of MHC class II-binding peptides is a more difficult classification problem than the prediction of class I molecules. The greater variability in length of MHC class II-binding peptides and their less well-characterized motifs make their alignment a difficult task, particularly as the vast majority contain more than one hydrophobic residue, allowing for multiple possible alignments. For instance it has been observed that application of a standard multiple alignment method, such as GCG Pileup (http://www.gcg.com/), failed to produce a useful alignment [1]. In these alignments a nonamer core was not preserved, nor would the sequences align relative to the primary anchors.

Anyhow, several methods have been used to predict MHC class II binding peptides, including those based on binding motifs, quantitative matrices and artificial neural networks (ANNs). Binding motifs specify which residues at given positions within the peptide are necessary or favorable for binding to a specific MHC molecule. Motifs for MHC class I molecules are relatively well defined. Nijman and co-workers [2] compared experimental results for binding to HLA-A2.1 with those obtained by motif-based prediction. Of 35 predicted binding peptides, they found that only 15 (43%) actually bound. With the exception of certain molecules [3], specific binding motifs for MHC class II molecules are less well defined [1]. Quantitative matrices are essentially refined binding motifs. They provide coefficients for each amino acid/position that can be used to calculate scores predictive of binding. The assumptions are that each residue contributes independently of other residues to binding and when located at a given position contributes the same amount to binding even within different sequences. Quantitative matrices have been defined for class II molecules [4]. ANNs are connectionist models commonly used for classification and pattern recognition tasks. ANNs used for the prediction of MHC class II binding peptides have achieved both positive and negative predictive values of nearly 80% [5]. Because of ambiguities resulting from the variable length of reported binders and the uncertain location of their core regions, peptides tested experimentally for binding and used as inputs to train an ANN require preprocessing by alignment relative to their binding anchors. An example is Mallios [6] who use an iterative stepwise discriminant analysis meta-algorithm to derive a quantitative motif for MHC class II. MHC class II-binding peptides also have more degenerate motifs. However, growing evidence supports the observation by Hammer et al. [3] that MHC class II-binding peptides contain a single primary anchor at the amino terminus, which is a hydrophobic amino acid (Y, F, W, I, V, L or M).

Each of the described prediction methods has its advantages and drawbacks. Binding motifs encode the most important rules of peptide/MHC interaction, but do not generalize well. Quantitative matrices can predict large subsets of binding peptides reasonably well, but cannot deal with non-linearity within data and may miss distinct subsets of binders. Also, quantitative matrices are not adaptive and self-learning, so that integration of new data usually requires redesigning of the matrix. ANNs can deal with non-linearity and are adaptive and self-learning, but require a large amount of preprocessed data. An ideal prediction method would integrate the strengths of these individual methods while minimizing their disadvantages. We have therefore developed SVMHCII based on the work of Dönnes and Elofsson [7], a hybrid method for the prediction of peptides that bind to MHC class II molecules. It utilizes: (i) local alignment for preprocessing; (ii) a Support Vector Machine to derive models; (iii) the utilization of 20 models as a *model cluster*. These clusters are constructed for the 26 MHC class II alleles whose binding peptides in the database MHCPEP are sufficient to train the SVM. Further, a comparison between a public available predictor, ProPred, and SVMHCII is made.

2 Background and Theory

2.1 Immune system

The immune system has evolved in vertebrates to protect them from invading pathogenic microorganisms and cancer. It is extremely adaptive and can generate a vast number of cells and molecules used to recognize and kill a limitless variety of foreign invaders. The immune system can be divided into the two interrelated parts, recognition and response. A foreign molecule can be distinguished from other foreign and self molecules by small chemical differences. Once the foreign molecule is recognized, an appropriate immune response can be raised.

2.1.1 MHC Cells involved in acquired immunity and specific recognition

There are three major cell types involved in acquired immunity. Two of these cells develop from a common ancestor and they are the B- and T lymphocytes. B cells mature in the bone marrow and T cells in the thymus. Antigen presenting cells (APC), such as macrophages and dendritic cells, are the third cell type.

The main feature of B and T cells is their specificity to antigen via antigen binding surface receptors. They are also responsible for other important features of immunology such as diversity, memory and self/non-self recognition. The antigen binding surface receptor of B lymphocytes is membrane bound antibodies. The T-cell receptor can only recognize antigen in conjugation with certain cell-membrane proteins known as the major histocompatibility complex (MHC) molecules. When a naive T-cell becomes activated by an antigen associated with a MHC molecule it differentiate into memory T-cells and various effector T cells.

The T Lymphocytes can be divided into two major groups, T-helper (Th) and T cytotoxic(Tc) cells. The distinction of the two subtypes is made by glycoproteins on their surface known as CD4 and CD8. T cells that have CD4 on their surface generally function as Th cells and those that have CD8 as Tc cells. An activated Tc cell may under the right stimulation, i.e. recognition of an antigen-MHC I molecule, proliferate and differentiate into a cytotoxic T lymphocyte (CTL). CTL's monitor the body for tumor cells and virus infected cells. They do not secrete much cytokines (immunological "communication" molecules), instead they can kill cells with their cytotoxic mediators. Th cells become activated when they interact with cells displaying MHC II molecules complexed with antigen. They secrete various cytokines as well as growth factors important in the activation of B cells.

APC on the other hand have no such antigen-specific receptors. Their function is to process and present antigens to specific T cell receptors (TCR).(The two functional molecules on APC used for antigen presentation are called MHC I and MHC II. The processed antigen is non-covalently bound to these molecules. The subset of T cells that are activated by MHC I molecules are the cytotoxic T cells. MHC II are present on APC and activate T helper cells.) The most important function of APC is to activate Th cells. Th cells are important in directing immune responses and therefore their activation must be carefully regulated. This activation of Th cells is as mentioned above carried out by antigen-MHC II molecules. The process is shown in figure 2.

2.2 Pattern recognition and machine learning

The term pattern recognition includes a wide range of information processing problems of great practical significance, from speech recognition and the classification of handwritten characters, to fault detection in machinery and medical diagnosis. [8]

2.2.1 Separable patterns

The best way to introduce separable patterns is to give a simple example. Let us start off with a hypothesis that men have bigger feet and weigh more than women. If this is true it should be possible to predict weather a person is a man or a woman, given the weight and shoe size. Table 1 shows weight, shoe size and gender for ten persons. The data in Table 1 is plotted in Figure 3, showing females marked '+' and men marked '*'. As can be seen in Figure 3 a line can be drawn to separate the data points into two groups, one for women and one for men. Since the plot is two dimensional there is a line separating the two groups. In a dimension of order N, the problem is to find a hyperplane in N-1 dimensions that separate the groups.

Figure 2: Degradation and transport of antigens that bind major histocompatibility complex (MHC) class II molecules. (a) In an antigen-presenting cell (APC), newly synthesized MHC class II molecules bind the invariant chain (IC), which prevents binding of peptides that are present in the endoplasmic reticulum (ER). (b) The IC allows transport of MHC class II molecules from the ER into the Golgi apparatus to acidified endosomes. (c) Endosomes contain peptides that are derived from either resident pathogens (e.g. bacteria) or (d) engulfed extracellular proteins (or pathogens) (e) in the phagosomes. (f) Proteases within the endosome degrade proteins into peptides. (g) The endosome fuses with the Golgi to form the trans-Golgi. (h) Here, the IC is cleaved and released from the MHC class II molecule. This allows the binding of peptides within the endosome to the peptide-binding cleft of the MHC molecules. An MHC-class-II-binding molecule (HLA-DM) binds to MHC class II molecules to facilitate the release of the IC (not shown). (i) The MHC class II-peptide complex is then transported to the cell surface of the APC for (j) recognition by the T-cell receptor (TCR) of (CD4+) T-helper lymphocytes (THLs) and (k) intracellular signaling for activation.

Figure 3: An example of weight and shoe size plotted for 10 persons. As can be seen in the figure, a line can be drawn to separate men from women

Weight	Shoe size	Gender
67	38	F
73	41	M
87	42	M
79	42	M
84	44	M
78	40	F
107	46	M
56	37	F
55	38	F
82	41	F

Table 1: Weight, shoe size and gender for 10 different persons.

2.2.2 Machine Learning

The term Machine Learning is generally used for automatic computing procedures based on logical or binary operations, that learn a task from a series of examples. When computers are used for solving practical problems, the required output from a given input can usually be described explicitly. A programmers task is here to set up a number of rules so that a given input gives the right output. The problem arises when very complex systems are to be analyzed. If no general rules are given it might be impossible to compute the desired output from a given input. An alternative is then to learn the input/output functionality from examples. An example of this is a child learning what cars are sports car simply by being told which of a large number of cars are sporty rather than by giving a precise specification of sportiness. This type of learning is called supervised learning and the examples of input/output are referred to as the training data [9].

The function that maps inputs to outputs is called the target function. The solution to the learning problem is an estimate of the target function and is the output of the learning algorithm. The quality of a learning algorithm can be assessed by the number of miss-classifications made during the learning phase. In a classification case the output is often referred to as the decision function. If a classification is of the type sick/healthy or regular car/sports car, it is called a binary classification. Multi-class classification deals with a finite number of classes.The ability of a function to map unseen data into the right class is known as generalization, and it is this property that one wishes to optimize. There are two main problems that make it hard for a learning algorithm to have good performance. One is that the function it tries to learn may not be easy to verify. The second problem is that the training data often is noisy and there is no guarantee that there exist a function that maps training data well.

By trying to optimize the generalization instead of the true function we have a more “loose” task to carry out. If our estimate of the true function gives the right output it satisfies the generalization criterion. We do not have any constraints on the size or “meaning” of our function now. The ability to generalize puts other constraints on the learning algorithm. We do not want a function that correctly maps all the training examples, but makes essentially uncorrelated predictions on unseen data. Functions like this are said to be over-fit. There are many different ways to keep away from over-fitting, e.g. keeping the complexity of the decision function low.

2.3 Support Vector Machine

Support vector machines are based on the Structural Risk Minimization principle from machine learning theory. The idea of structural risk minimization is to find a hypothesis H for which we can guarantee the lowest true error. The true error of H is the probability that H will make an error on an unseen and randomly selected test example. An upper bound can be used to connect the true error of a hypothesis H with the error of H on the training set and the complexity of H the hypothesis space containing H. Support vector machines and the hypothesis H which (approximately) minimizes this bound on the true error by reactively and efficiently controlling the VC-Dimension of H. SVMs are very universal learners. In their basic form, SVMs learn linear threshold functions. Nevertheless, by a simple plug-in of an appropriate kernel function, they can be used to learn polynomial classifier, radial basic function (RBF) networks, and three-layer sigmoid neural nets. One remarkable property of SVMs is that their ability to learn can be independent of the dimensionality of the feature space. This means that even in the presence of very many features, if our data is separable using functions from a lower dimensional hypothesis space. The same margin argument also suggests a heuristic for selecting good parameter settings for the learner. The best parameter setting is the one which produces the hypothesis with the lowest dimension. This allows fully automatic parameter tuning without the expensive cross-validation, that is necessary in Artificial Neural Networks.

2.4 Cross-validation

It is important to put a measure on the whole procedure of SVM learning and classification. It is also important to put some relevant statistics on the performance measurement. Even if SVM's are less prune to over-training than other machine learning methods we need to test the performance on target data that is "unseen", i.e. a model should be established without target data training. By measuring the ability of the model to predict the target the final performance can be calculated. How to choose the target and how many targets should be used are two common questions to be answered. The method of cross-validation is used in this project. In cases where the amount of labeled data is limited, cross-validation can be used. The idea of cross-validation is to split the training set at random into N subsets. N minus one sets are then used for training and the remaining set is used for testing the performance. The procedure is then repeated for all the N subsets. One drawback of this approach is that the training procedure must be repeated N times and each one might need a lot of computer time.

2.5 Performance measurements

The most obvious measurement would be to calculate the fraction correct predictions. However, in our case only a small fraction of all peptides binds and therefore the predictions that no peptides binds to the MHC molecule would be very good using this measure, i.e. we need to use other measurements described below.

In two-class cases where the output from a prediction algorithm is continuous, the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) depend on where a threshold is drawn. When some of the data is used as test-set, they are predicted as either binding or non-binding. A true binder that is predicted to be a binder is called a true positive

(TP), a true binder that is predicted to be a non-binder is called a false negative (FN), a true non-binder predicted to be a non-binder is a true negative (TN) and a true non-binder predicted to be a binder is a FP. These test examples are taken through the SVM and are predicted to be either a binder or a non-binder. In general there is a trade-off between the amount of false positives and false negatives produced by the algorithm. One way to summarize this is ROC (receiver operating characteristics). A ROC plot displays for different thresholds the sensitivity (TP/(TP+FN)) versus false positive rate (FP/(FP+TN)). Another possibility is to plot the sensitivity against specificity (TP/(TP+FP)) in a similar plot.

The four different types of hits can also be used to calculate the Matthews Correlation coefficient(Mc):

$Mc = \frac{(TP \cdot TN)-(FP \cdot FN)}{\sqrt{(TN+FN)(TN+FP)(TP+FN)(TP+FP)}}$

The Matthews correlation can vary between -1 and 1. A value of 1 means a perfect prediction, 0 equals a prediction no better than random and -1 equals total opposite predictions.

3 Material and Methods

3.1 MHC peptide databases

3.1.1 MHCPEP: MHC binding database

MHCPEP [1] [10] is a curated database comprising over 13 000 peptide sequences known to bind MHC molecules. Entries are compiled from published reports as well as from direct submissions of experimental data. Each entry contains the peptide sequence, its MHC specificity and where available, experimental method, observed activity, binding affinity, source protein and anchor positions, as well as publication references. The binding peptides of MHC II are extracted from MHCPEP database using a Perl script.

3.1.2 ENSEMBLE: non-binding database

The Ensembl[2] database project provides a bioinformatics framework to organize biology around the sequences of large genomes. It is a comprehensive source of stable automatic annotation of the human genome sequence, with confirmed gene predictions that have been integrated with external data sources. Since less than one percent of protein sequences are expected to bind to MHC class II the Ensembl database can be considered as a non-binding peptide database. All the non-binding peptides are extract from this database, using a Perl script.

3.2 SVM^{light}

The SVM software used in MHC class II binding prediction is SVM^{light} by Thorsten Joachims (Version: 4.00 Release Date: 11.02.2002).[3] SVM-light is an implementation of SVM for the problem of pattern recognition. The optimization algorithm used in SVM^{light} is described in Joachims [11]. The algorithm has scalable memory requirements and can handle problems with many thousands of support vectors efficiently. There is a set of kernels, or transformation functions, available for the construction of the support vectors.

[1] http://wehih.wehi.edu.au/mhcpep/

[2] http://www.ensembl.org

[3] http://svmlight.joachims.org/

3.2.1 SVM_learn and SVM_classify

When the SVM$^{\text{light}}$ source code is downloaded and compiled, it gives two ”functional” modules for the users. SVM_learn is used to train a model and the input is the input/output pairs of our training examples. In our case with a 9 AA peptide sequence the input vectors are 181 elements long (i.e. 180 elements for the 9 AA and an additional 1 or 0 for binder/non-binder). When a model is learned by the SVM_learn module it can be used for classification by the SVM_classify module.

3.2.2 Parameters in SVM$^{\text{light}}$

For SVM$^{\text{light}}$ there are several parameters to handle the model, the most important is the kernel type. The most common kernels in SVM are linear, polynomial, and radial basis function (RBF). The kernels definition is shown in the table 2. Examples of other parameters that may be changed are: trade-off between training error and margin, cost-factor-by which training errors on positive examples out-weight errors on negative examples and different kernel specific parameters. There is no exact theory of how to choose parameters. The choice of parameters that give the optimum classification has to be investigated for each functional class and is of central importance to obtain a good model. The procedure of choosing parameters is carried out by using a nested loop, i.e systematic searching for the best combination of all parameters. Training is carried out using three-fold cross validation.

Code	Name	Definition
0	linear	$K(x_i, x_j) = sx_ix_j + c$
1	polynomial	$K(x_i, x_j) = (sx_ix_j + c)^d$
2	rbf	$K(x_i, x_j) = exp(-\frac{\|\|x_i - x_j\|\|^2}{\delta^2})$

Table 2: List of Kernel in SVM

3.3 Algorithm for prediction

The process in the algorithm includes: preprocessing of the binding database, initialization, *model cluster* establishment and classification. An overview of the algorithm is shown in figure 4.

Figure 4: The overall algorithm steps for SVMHCII.

3.4 Preprocessing the database

Preprocessing the binding peptide is necessary. Firstly, preprocessing can avoid the over-training of the data. If two or more same or similar sequences are used to train the model, then the model might have a bias to these. Secondly, it is not clear which part of the sequences really binds to the receptor. Preprocessing can be used to calculate the most likely binding region. The calculation is based on the following two ideas: the binding part is 9 AA long and if two subsequences share identical subsequence of length 9 or longer, then the binding part exists in this subsequence. MHCPEP contain many peptides binding to the same allele with similar long common subsequences. For example, for allele HLA-DR51(DRB5*0101) four peptides (PVVHFFKNIVTPRTPPY, VVHFFKNIVTPRTPPY, VHFFKNIVTPRTPPY and HFFKNIVTPRTPPY) are in the database. Most likely they all use the same binding region, the peptide HFFKNIVTPRTPPY. To detect similar subsequences local alignments between all peptides binding to the same allele are performed. If two sequence's share more than more than 7 identical residues the shorter of them is excluded from further studies. By this method, a reduced database that contains no peptides with identical parts is created. The amount of data for each allele in MHCPEP and the reduced database is shown in table 3.

No	Allele name	MHCPEP	Reduced Database
1	HLA-DR4(DRB1*0401)	504	197
2	HLA-DR1	460	142
3	HLA-DR1(DRB1*0101)	281	111
4	HLA-DR11(DRB1*1101)	170	81
5	HLA-DR2	296	76
6	HLA-DQ4(DQA1*0302xDQB1*0401)	101	69
7	HLA-DR7(DRB1*0701)	149	57
8	HLA-DR5	199	53
9	HLA-DQ7(DQB1*0301)	165	46
10	HLA-DR4(DRB1*0402)	139	46
11	HLA-DR3	115	45
12	HLA-DR7	211	43
13	HLA-DR8	59	40
14	HLA-DR4	117	38
15	HLA-DR4(DRB1*0404)	93	37
16	HLA-DR8(DRB1*0801)	66	36
17	HLA-DQ8(DQA1*0301xDQB1*0302)	98	35
18	HLA-DR4(DRB1*0405)	152	34
19	HLA-DR51(DRB5*0101)	124	31
20	HLA-DR52(DRB3*0101)	44	27
21	HLA-DR3(DRB1*0301)	77	24
22	HLA-DR15(DRB1*1501)	102	23
23	HLA-DR17(DRB1*0301)	44	23
24	HLA-DR17	117	22
25	HLA-DQ2	24	21
26	HLA-DR9(DRB1*0901)	71	20

Table 3: The comparison of MHCPEP and the reduced database

3.4.1 Initialization

The aim of our initialization procedure is to make an initial guess of binding regions that can be used in the step-wise training described below. Initially a binding nonamer for every peptide in the reduced database is generated. Here, all nonamers of a peptide are assumed to contain the binding region. A SVM prediction model is made for these assumed binding nonamers.

3.4.2 Model establishment and cluster

The *model clusters* are made using a “stepwise” algorithm, shown in figure 5. At step 0, the initial model is used for the binding and non-binding peptides using 4-fold cross validation. For each peptide the subsequence with the highest predicted binding is predicted. Those subsequence then enter the binding dataset for the next step. At the same time, all other subsequences of this peptide are excluded. Using this new set of binding data a new SVM is trained using the current binding dataset. Using these predictions a new “binding” dataset is generated. This set is then used to train new SVMs, that are used to predict new binding regions etc. This process continues for 23 rounds.

Figure 5: The algorithm of *model cluster* establishment

3.4.3 The prediction using *model clusters*

The query peptide will be separated into 9 AA subsequences,and all the 9 AA subsequence can be classified by the models in the cluster. If a subsequence is recognized a score of one will be added. After the classification, subsequence with a score higher than the cutoff will be suggested as a binding peptide. For the entry HFFKNIVTPRTPPY (14 AA), 6 subsequences are generated HFFKNIVTP, FFKNIVTPR, FKNIVTPRT, KNIVTPRTP, NIVTPRTPP, IVTPRTPPY. After the classification by the cluster for HLA-DRB5(0101), the result is shown in table 4. HFFKNIVTP get a score of 20, which means twenty cluster models recognize it as binding region.

Rank	Sequence	Start	Stop	Score
1	HFFKNIVTP	1	9	20
2	FFKNIVTPR	2	10	0
3	FKNIVTPRT	3	11	0
4	KNIVTPRTP	4	12	0
5	NIVTPRTPP	5	13	0
6	IVTPRTPPY	6	14	0

Table 4: The result of one *model cluster* prediction.

3.5 Comparison with ProPred

ProPred is a free public MHC class II binding prediction server.[4] [4] This server uses quantitative matrices derived from published literature by [12] and also MHCPEP. To test our *model cluster* methods, a comparison is made between SVMHCII and ProPred. Eleven different alleles can be predicted by both servers. For each of then 50 binding peptides are extracted from MHCPEP and 200 non-binding peptides are then selected randomly. The threshold for ProPred is set at 3% (default), while the cutoff for the SVMHCII was set to be 5 after optimization. A perl script is used to generate test peptides and visit both servers. The results from the comparison are studied using the same measures as described above.

3.6 Test-set validation

To test the ability of the *model cluster* to recognize new MHC binding peptides MHC molecule, a second cross validation test was also used. A quarter of the peptides from the reduced database were selected as target. Because none of the peptides in the reduced database are similar to each other these target should be unseen from the training data. Non-binding peptides were selected from Ensembl with a variable length from 9 to 16 residues. *Model clusters* were made on the

[4] http://www.imtech.res.in/raghava/propred/index.html

remaining binding peptides and ten times as many non-binding 9-AA peptides. This process was then repeated five times to test the average ability to predict new peptides. In total 13 alleles with more than 40 peptides in the reduced database were tested. Here we found that the cutoff of one, i.e. that the test peptides is recognized by one model of the cluster, was best. Finally the accuracy and Mc coefficient were calculated.

4 Results and Discussion

4.1 Results from Model clusters

For every step, a Mc coefficient is calculated during the cross-validation. The Mc for HLA-DRB5*0101 is calculated and recorded for 23 iterations, see figure 4.1. First, it can be observed the the Matthews correlation coefficients are higher than zero, i.e. all prediction are better than random. Figure 4.1 indicates that during the first few steps, the Mc improves, and then fluctuates between 0.6 and 0.8. For this reason, models made during the first 3 steps will be skipped in the cluster establishment, i.e. models made in step 4 to step 23 enter the clusters. For a general evaluation of the performance for each allele, the average Mc during the 20 steps of model evaluation is shown in table 5. For the 26 alleles, the average Mc vary from 0.57 (HLA-DQ2) to 0.78 (HLA-DR3).

4.2 Comparison with ProPred

Because ProPred only can predict peptides longer than 9 AA, all the non-binding peptides are set at length between 10 to 15 AA, i.e. all nonamers in the target database are ignored. Table 6 shows the result of comparison between SVMHCII and ProPred. Only in one case of 11 alleles, the Mc of ProPred prediction is better than SVMHCII. Figure 8 shows the overall comparison. Figure 7 shows four examples of a sensitivity-specificity plot between the predictions of SVMHCII and ProPred. From the plots we can see that SVMHCII performs rather well. Even for the allele HLA-DR15(DRB1*1501), which is the only allele in the Mc comparison when ProPred is better, the for which Se-Sp plot of SVMHCII performs better than that of ProPred.

Figure 6: The Mc in cross-validation for 23 steps

No	Allele name	Average Mc
1	HLA-DR17(DRB1*0301)	0.67
2	HLA-DR7	0.66
3	HLA-DR4(DRB1*0402)	0.67
4	HLA-DR51(DRB5*0101)	0.74
5	HLA-DR15(DRB1*1501)	0.74
6	HLA-DR52(DRB3*0101)	0.69
7	HLA-DR5	0.68
8	HLA-DR11(DRB1*1101)	0.69
9	HLA-DR4	0.59
10	HLA-DQ4(DQA1*0302xDQB1*0401)	0.73
11	HLA-DQ7(DQB1*0301)	0.70
12	HLA-DQ8(DQA1*0301xDQB1*0302)	0.63
13	HLA-DR3	0.78
14	HLA-DR3(DRB1*0301)	0.68
15	HLA-DR2	0.73
16	HLA-DR4(DRB1*0404)	0.66
17	HLA-DR8(DRB1*0801)	0.65
18	HLA-DR1(DRB1*0101)	0.67
19	HLA-DR7(DRB1*0701)	0.68
20	HLA-DR8	0.67
21	HLA-DQ2	0.57
22	HLA-DR9(DRB1*0901)	0.65
23	HLA-DR4(DRB1*0401)	0.62
24	HLA-DR1	0.60
25	HLA-DR4(DRB1*0405)	0.58
26	HLA-DR17	0.66

Table 5: The average Mc for each allele in the 20 times cross-validation

4.3 Test-set validation

The above results indicate that the *model clusters* predict binding peptides from the current database quite well. However, we also want to examine the predictions for unseen peptides. As an example we used the allele HLA-DR11(DRB1*1101), see table 4.3. For the give different sets of test-set predictions for the cluster, the average Mc is 0.62, i.e. only slightly lower than the Mc obtained in the cluster establishment (0.69). This indicated the ability of the clusters to recognize unseen peptides is acceptable.

Figure 7: Sensitivity-Specificity plot of four alleles examples between SVMHCII and ProPred

5 Conclusions and future steps

5.1 Conclusions

In this thesis, a new Major Histocompatibility Complex class II binding approach, SVMHCII, is developed. For 26 different MHC-alleles *model clusters* were created. Each *model cluster* contains 20 SVM based models made using a step-wise algorithm. Our results indicate that our method perform at least on par with alternative methods.

The use of *model cluster* instead of one single model is better. Since the real binding region of the MHC-II binding peptides is unknown it is difficult to train the model using machine learning methods. The model cluster approach provided a method to solve this problem. All possible binding regions (assumed to be monomeric subsequences in this project) have a chance to enter the *model cluster* in some step. We have reasons to believe that the most frequently appearing peptides are

Allele Name	Mc Pro	MC SVM	Sp Pro	Sp SVM	Se Pro	Se SVM
HLA-DR4(DRB1*0401)	0.48	0.58	0.59	0.55	0.54	0.84
HLA-DR1(DRB1*0101)	0.48	0.79	0.60	0.72	0.56	0.98
HLA-DR11(DRB1*1101)	0.35	0.73	0.36	0.70	0.60	0.90
HLA-DR7(DRB1*0701)	0.21	0.64	0.37	0.55	0.38	0.94
HLA-DR4(DRB1*0402)	0.31	0.63	0.47	0.53	0.42	0.96
HLA-DR4(DRB1*0404)	0.39	0.66	0.53	0.54	0.48	1.00
HLA-DR8(DRB1*0801)	0.46	0.76	0.54	0.65	0.62	1.00
HLA-DR4(DRB1*0405)	0.18	0.64	0.38	0.59	0.27	0.88
HLA-DR51(DRB5*0101)	0.27	0.62	0.45	0.56	0.34	0.9
HLA-DR15(DRB1*1501)	0.57	0.49	0.71	0.47	0.60	0.8
HLA-DR3(DRB1*0301)	0.28	0.69	0.50	0.64	0.3	0.9

Table 6: Comparison of the Mc coefficients (Mc), Specificity (Sp), and Sensitivity (Se) for twelve MHC class-II alleles between SVMHCII (SVM) and ProPred (pro)

Figure 8: A comparison between ProPred and SVMHCII. For each cases, a star mark will be plot for the corresponding Mc.

the most likely binding regions. Therefore a *model cluster* can provide more stable results than a single model. The average Mc coefficients are 0.57-0.74, better than reported in earlier studies. The clusters can provide varies of predictions with different sensitivity and specificity. By changing the cutoff, the user can easily choose different sensitivity and specificity. Normally, an unseen binding peptide should be recognized by at least one model in the cluster and if a peptide is determined by more the 15 models, it is almost certain that it is binding.

Times	TP	TN	FP	FN	Sp	Se	Mc
1	13	76	5	8	0.72	0.62	0.59
2	10	75	6	11	0.63	0.47	0.44
3	15	76	5	6	0.75	0.71	0.67
4	16	76	5	5	0.76	0.76	0.70
5	12	80	1	9	0.92	0.57	0.68
Overall	66	383	22	39	0.75	0.63	0.62

Table 7: The Results of HLA-DR11(DRB1*1101) validation

5.2 Future development

For MHC class II alleles experimental methods to define the actual binding region of a peptide for every peptides would be very useful. Though most of the peptides in Ensembl are non-MHC II-binding. A database containing experimentally verified non-binding protein would also be very useful. Further more data would be needed to expand the predictions to more alleles. Also more new data would be useful to test the final models.

6 Acknowledgments

I would like to thank my supervisor Arne Elofsson at Stockholm Bioinformatics Center for his help and valuable ideas throughout the project. I also want to thank Bjorn, Isaac, and all the staffs at Stockholm Bioinformatics Center for their help and support during this project. I'd also like to thank Pierre Dönnes. Though I even haven't seen him so far. His work in MHC class I predictions really helped me. Specially, I'd like to appreciate my aunt, my uncle and Feifei for supporting my life in Sweden. Finally, I thank my family and friends for their doubtless support and for them spurring me to fresh efforts.

7 References

References

[1] Rammensee H-G. Identification of T-cell epitopes using allele-specific ligand motifs. Behring Inst Mitt. 1994 Dec;(95):7-13

[2] Nijman HW, Houbiers JG, Vierboom MP, van der Burg SH, Drijfhout JW, D'Amaro J, Kenemans P, Identification of peptide sequences that potentially trigger HLA-A2.1-restricted cytotoxic T lymphocytes. Eur J Immunol. 1993 Jun;23(6):1215-9.

[3] Hammer J, Bono E, Gallazzi F, Belunis C, Nagy Z, Sinigaglia F. Precise prediction of major histocompatibility complex class II-peptide interaction based on peptide side chain scanning. J Exp Med. 1994 Dec 1;180(6):2353-8.

[4] Singh H, Raghava GP. ProPred: Prediction of HLA-DR binding sites. Bioinformatics, 2001 17(12), 1236-37.

[5] Brusic V, Rudy G, Honeyman G, Hammer J, Harrison L.Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network.Bioinformatics. 1998;14(2):121-30.

[6] Mallios RR, Predicting class II MHC/peptide multi-level binding with an iterative stepwise discriminant analysis meta-algorithm. Bioinformatics. 2001;17(10):942-948.

[7] Dönnes P, Elofsson A. Prediction of MHC Class I binding Peptides, using SVMHC, BMC Bioinformatics 2002; 1471-2105/3/25

[8] Bishop C.M., Neural Networks for Pattern Recognition, Clarendon Press, Oxford,1995

[9] Cristianin N., Shawe-Taylor J., Support Vector Machines and Other Kernel-Based Learning Methods, University Press, Cambridge, 2000

[10] Brusic V, Rudy G, Kyne AP, Harrison LC. MHCPEP, a database of MHC-binding peptides: update 1997 Nucleic Acids Res. 1998 26(1):368-71.

[11] Joachims T, 11 in: Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning, B. Schölkopf and C. Burges and A. Smola (ed.), MIT Press.

[12] Sturniolo T, Bono E, Ding J, Raddrizzani L, Tuereci O, Sahin U, Braxenthaler M, Gallazzi F, Protti MP, Sinigaglia F, Hammer J., Generation of tissue-specific and promiscuous HLA ligand databases using DNA chips and virtual HLA class II matrices. 1999 Nature Biotechnology 17, 555-562.