CS229 Final Project Report

Predicting Epitopes for MHC Molecules

Xueheng Zhao, Shanshan Tuo

Biomedical informatics program
Stanford University

Abstract

Major Histocompatibility Complex (MHC) plays a key role in immune response by presenting antigenic peptides, which are recognizable to T-cells. Identifying MHC-binding peptides is crucial to understand pathogenesis and develop corresponding vaccines. Direct identification of MHC-binding peptides by biological assays is laborious and expensive, because of the huge size (20^9) of potential combinations. Current computational methods are also not satisfactory: many of them failed to capture the features of non-conserved motifs, MHC molecule polymorphism, and non-specific binding to low-affinity peptides. Large datasets from National Institute of Health (NIH) public data repository and our own research data, with machine learning methods, provide a great opportunity for addressing this problem. Thus, we aim to develop an efficient and accurate method to identify and predict MHC-binding peptides and to further differentiate various MHC subtypes by their peptide-binding specificities. By selecting ~3000 peptide motifs as features, we built classifiers with Naive Bayes and SVM-based approaches. These classifiers achieved accuracy up to 99% on four most frequent human MHC subtypes—HLA-A01, HLA-A02, HLA-B27, and HLA-B08—with 10 fold cross-validation. We applied these classifiers onto experimental data—a pool of potential binders. It has been found that up to 98% peptides are classified as binders and the classifier can be used to determine the specific subtypes. We also applied another approach, iterative statistical search in feature optimization. This iterative approach showed very promising results in our preliminary test. By applying this approach in feature generation, we achieved about 85% accuracy with less than 30 features. Further exploration of this approach will optimize feature space and improve algorithm efficiency.

Introduction

The immune system acts as a physical barrier against pathogen infections by activating B cells and T cells. B cells and T cells are special types of white blood cells that can recognize "nonself" cells, including pathogen-infected cells, and trigger immune response. In particular, cytotoxic T cell receptors can bind to major histocompatibility complex (MHC) that are antigen-specific receptors on the "nonself" cells and alert immune system to kill these infected cells (Smith-Garvin et al., 2009). Among three classes (I, II, and III) of MHC family, class I is by far the most well characterized subgroup, which we focused on in this study. Class I MHC proteins have a special structure with four antiparallel β-strands in the center region and two α-helices on one side. The two α-helices form a groove that contains six amino-acid binding pockets and can only accommodate short peptides of 8 to 11 amino acids—short fragments of antigens, in other words epitopes(Figure 1; Mester et al., 2011). Identifying and predicting MHC-binding epitopes are essential to understand the cause of diseases and develop corresponding vaccines (Lundegaard et al., 2007).

Due to the size of the potential binding peptides (20^9 = 512 billion) for each MHC molecule (Liao and Arthur, 2011), empirical approach by biology experiments is laborious

and expensive to identify binding peptides. Multiple computational approaches including artificial neutral network (ANN) based method were developed to predict binding peptides (Lundegaard, 2008). But many of them failed to capture and quantify the complex MHC-binding properties due to few conserved peptide motifs, polymorphism in MHC molecules, and binding ability to low-affinity peptides (Nielsen et al., 2004; Peters et al., 2005). To overcome these challenges and better predict the epitopes from binding peptides of different MHC molecules, we studied the characteristics of the peptides and MHC protein sequences near the peptide-binding region. We extracted more than 12,000 peptides that bind to four most frequent MHC subtypes—HLA-A01, HLA-A02, HLA-B27, and HLA-B08 from Immune Epitope Database 2.0 and resource analysis (IEDB). We then selected and optimized features to built classifiers. Using machine-learning algorithms, we are able to predict binding peptides from a lab-generated dataset.

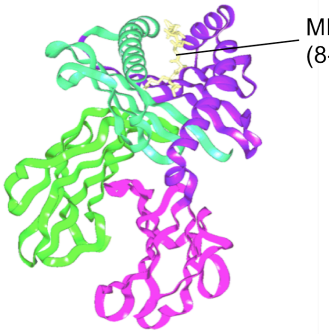

Figure 1 Structure and binding groove of major histocompatibility complex (MHC) molecules (Adapted from website of Institute of Structural and Molecular Biology at University of London)

Data collection

Binding-peptides were collected from Immune Epitope Database 2.0 and resource analysis (IEDB), and newly generated data from our lab at systems biology department by high-resolution mass spectrometry. From IEDB database, we extracted four datasets of MHC-binding peptides to four MHC subtypes—HLA-A1, HLA-A2, HLA-B27, and HLA-B08—each dataset containing about 3000 binding and 3000 nonbinding peptides. For lab-generated MHC-binding peptides, we purified cell lysate proteins of B cells that associate with MHC proteins and sequenced the short MHC-binding peptides with orbitrap mass spectrometry. We focused on 9-mer peptides, because it is the dominant form that binds to MHC class I alleles. Our lab data include 786 distinct peptides, which are a pool of potential binding peptides to all four MHC subtypes.

To generate MHC-non-binding peptides, we applied Markov Model based on Uniprot protein database (Elias, 2012). The probability that these peptides actually bind MHC is really low and can be ignored.

Feature Selection

Motifs—short conserved sequence patterns of amino acids or nucleotides—are usually associated with key functional sites involved in catalysis and/or binding to other molecules. Considering that MHC molecules have preference to some amino acids at certain positions, we selected motifs as features to build the classifier for epitope binding prediction. We define a motif in our analysis as a 9-amino acid over an alphabet of 20 amino acids {ACDEFGHIJKLMNPQRSTVY} and '.' denoting any of the 20 possible amino acids. A sequence $s = s_1 s_2 \ldots s_9$ is said to contain a certain motif if it contains exactly the same amino acid patterns.

Protein sequence motifs are typically extracted from non-gapped regions (blocks) of a multiple sequence alignment (Figure 2). Each position in the motif represents the variability in a column of the block. In our study, features were selected as follows: for each position, select amino acids that have at least 5% overall occurrence in the dataset; build features containing 1 or 2 selected amino acid (Figure 2). Total number of features varies somewhat between different datasets because of different occurrence rate for individual amino acid.

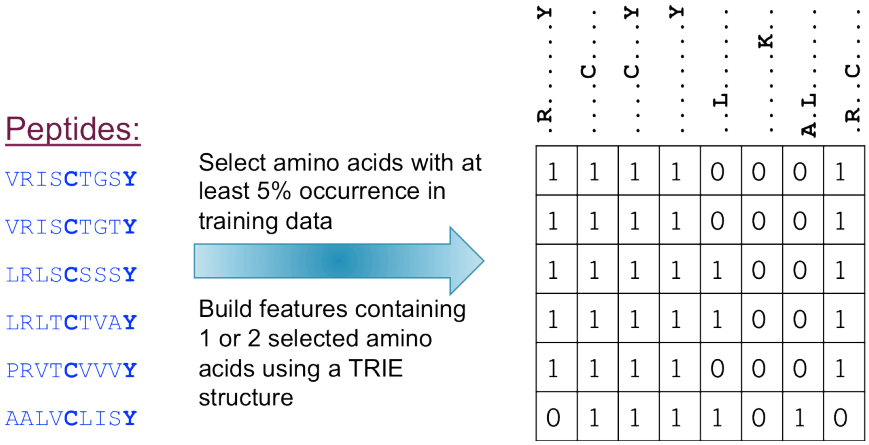

Figure 2. Feature selection for machine learning approaches

Machine learning approach

To develop the classifier for MHC-binding peptides, we used machine learning algorithms including Naive Bayes and support vector machine (SVM). In both methods, we pre-processed four groups of raw data (one for each MHC subtype) with feature mapping: each 9-mer peptide was transformed and represented by a high dimensional vector based on features selected as described above.

The processed data were fed into Weka to build models with Naive Bayes and SVM. 10-fold cross validation using the binding and nonbinding datasets was performed to compare methods of Naive Bayes and SVM with different kernel functions. Accuracies and ROC areas were obtained from Weka reports and plotted for visualization.

We tested our four SVM models (one for each MHC subtype) to predict the binding peptides to each MHC subtype from experimental peptide dataset.

Iterative statistical search method

To find significant motifs and decrease search space to improve algorithm efficiency, we developed a python program to statistically search the significant motifs from a peptide dataset based on Schwartz and Gygi (2005). Two frequency matrices (20 by 9 in dimension) of all residues at every position were built from binding and nonbinding datasets collected as above. Each row represents an amino acid; each column represents the position; value at row i and col j represents the probability of observing residue i at position j given a probability p, which was calculated for this residue/position pair from the background data set. A greedy recursive algorithm and pruning procedure were applied to search the sequence space to identify highly correlated residue/position pairs with significant binomial probability values. Thus, the sequence database is decomposed into a list of significant motifs and we can apply these motifs as features in our learning methods.

Results and Discussion

SVM-based method with ~3000 features gave high accuracy

To obtain the optimal features, we compared the overall accuracy with different number of features in 10-fold cross validation of Naive Bayes and SVM, i.e. motifs includes one or two frequent amino acids and motifs with one, two and three frequent amino acids, and with results from statistical iterative search (Table 1). Our preliminary results showed that SVM-based method with about 3000 features achieved accuracy close to 99%. Increasing features to 4,5000 motifs would not contribute much for accuracy but rather wastes memory space and increases running time.

Table 1. Overall accuracy by different feature selections

	Motif sets with 1 or 2 amino acids (~3000)	Motif sets with 1 to 3 amino acids (~45,000)	Statistical iterative search generated features (~25)
Naïve Bayes	72.0%	NA	85.9%
Support Vector Machine (SVM)	99.0%	99.8%	88.7%

So we decided to restrain our feature space to 3000 in the following study.

For each dataset for a particular MHC subtype, we compared the results from 10-fold cross validation (Figure 3). For all MHC subtypes, SVM provided a high accuracy and area under a ROC curve close to 1. On the other hand, Naive Bayes algorithm gave an area around 0.7 for subtypes HLA-A02, HLA-B27, HLA-B08, but achieved high value close to 1 for HLA-A01 subtype. Overall Naive Bayes achieved inferior performance comparing with SVM based method.

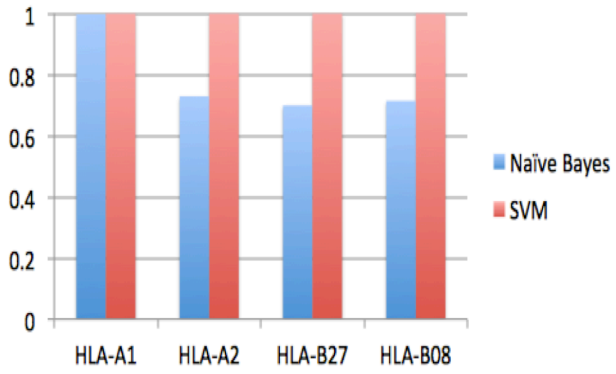

Figure 3. ROC area in 10-fold cross validation using Naive Bayes and SVM for four different MHC subtypes

MHC subtypes differentiated by peptide-binding specificity

By applying our trained SVM based classifier on our experimental data, we were able to determine which peptides MHC subtypes can specifically bind to (Figure 4). It is a little unexpected that HLA-A02, HLA-B08 and HLA-B27 subtype binds to a group of peptides that have no affinity to HLA-A01. The structural difference among these subtypes needs to be investigated to fully understand the underlying mechanism affecting binding affinity.

Statistical iterative approach gave relatively high accuracy with small feature space

To further optimize features for a better efficiency and specificity, we adopted a statistical iterative search method to generate relevant motifs from training dataset. Since only significant motif can be generated from this search, we are able to decrease our features to a less than 30 motif set. Using these features to train our SVM based classifier and applied on the prediction of our lab-generated dataset, we achieved 89% positive identification. This result showed that the patterns identified through an iterative search would be useful in the optimization of feature selection. Due to the time limit for this project, we were unable to pursue further in optimizing feature selection and applying this method to binding prediction.

Validation of algorithm

To evaluate the performance of our algorithm, we compared the performance of our approaches with NetMHC epitope webserver 3.2. NetMHC-3.2 used artificial neural network training method, which has been benchmarked as the best among available methods. It achieved about 75% confirmed accuracy on a large set of pathogenic viral proteome (Lundegaard et al. 2008). Using our experimental data, NetMHC predicted 64% as binders for four MHC subtypes, whereas our SVM-based method achieved 89% positive

prediction. The reason that our approach showed much better performance might due to our using the most updated IEDB dataset which covering more MHC alleles thus our approach picks up more features overlooked by NetMHC.

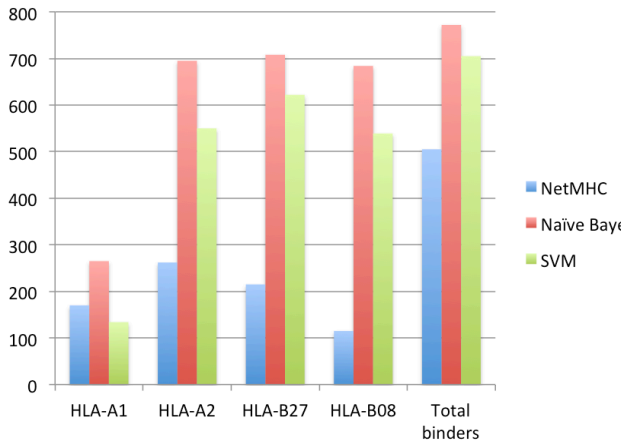

Figure 4. Binders identified by NetMHC, Naive Bayes, and SVM methods from lab-generated data

Future work

In this study, we identified better features to represent MHC properties and applied and compared Naive Bayes and SVM-based methods to predict and identify MHC-binding peptides and to further differentiate MHC subtypes.

In the future, we plan to scale up our data sets by importing more peptides from databases and test our developed method; optimize feature selection by integrating the flanking region motifs with the binding peptide motifs; improve motif identification and subtype identification algorithm by incorporating more MHC structural and functional properties combined with different kernel functions.

References

Elias, J.E. Markerofpeptide Perl script for peptide generation, 2012

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11, 1, 2009

Institute of Structural and Molecular Biology (ISMB), University of London http://www.cryst.bbk.ac.uk/pps97/assignments/projects/coadwell/004.htm

Liao, W., Arthur, J. Predicting peptide binding to major histocompatibility complex molecules. *Autoimmunity Reviews* 10: 469-473, 2011

Lundegaard, C., Lund, O., Kesmir, C., Brunak, S., and Nielsen, M. Modeling the adaptive immune system: predictions and simulations. *Bioinformatics*, 23, 3265-3275, 2007

Lundegaard, C., Lamberth, K., Harndahl, M., Buus, S., Lund, O., and Nielsen, M. NetMHC-3.0: accurate web accessible predictions of human, mouse, and monkey MHC class I affinities for peptides of length 8-11 *Nucleic Acids Research*, 36, W509-W512, 2008

Mester, G., Hoffmann, V., Stevanoic, S. Insights into MHC class I antigen processing gained from large-scale analysis of class I ligands. *Cellular and Molecular Life Sciences* 68:1521–1532, 2011

Nielsen, M., Lundegaard, C., Worning, P., Hvid, CS., Lamberth K et al. Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach. *Bioinformatics*, 20:1388-1397, 2004

Peters, B., Sette, A. Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method. *BMC Bioinformatics*, 6:132, 2005

Smith-Garvin, J., Koretzky, G., Jordan, M. T cell activation. *Annual Review of Immunology* 27: 591-619, 2009.

Schwartz, D., and Gygi, S. An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. *Nature Biotechnology* 23: 1391-1398, 2005

Vita, R., Zarebski, L., Greenbaum, J.A., Emami, H., Hoof, I. et al. The immune epitope database 2.0. *Nucleic Acids Res*. 38(Database issue):D854-62, 2010