PREDICTING THE BINDING AFFINITY OF MHC CLASS II PEPTIDES

Fatih Altiparmak[1], Altuna Akalin[2], Hakan Ferhatosmanoglu[1]
[1]Computer Science and Engineering, The Ohio State University
[2]Computational Biology Unit, Bergen Center for Computational Science, University of Bergen
emails: {altiparm, hakan}@cse.ohio-state.edu, Altuna.Akalin@bccs.uib.no

MHC (Major Histocompatibility Complex) proteins are categorized under the heterodimeric integral membrane proteins. The MHC molecules are divided into 2 subclasses, class I and class II. Two classes differ from each other in size of their binding pockets. Predicting the affinity of these peptides is important for vaccine design. It is also vital for understanding the roles of immune system in various diseases. Due to the variability of the locations of the class II peptide binding cores, predicting the affinity of these peptides is difficult. In this paper, we proposed a new method for predicting the affinity of the MHC Class II binding peptides based on their sequences. Our method classifies peptides as binding and non-binding. Our prediction method is based on a 3-step algorithm. In the first step we identify the informative n-grams based on their frequencies for both classes. In the next step, the alphabet size is reduced. At the last step, by utilizing the informative n-grams, the class of a given sequence is predicted. We have tested our method on the MHC Bench IV-b data set [13], and compared with various other methods in the literature.

1. INTRODUCTION

MHC (Major Histocompatibility Complex) proteins are categorized under the heterodimeric integral membrane proteins. The primary function of MHC proteins is presentation of antigenic peptides, which are degraded from foreign proteins, to T lymphocytes so that an immune response in the system can start [4]. The MHC molecules are divided into two subclasses, class I and class II. Members of both classes bind to peptides by recognizing a core sequence having 9 residues. Two classes differ from each other in size of their binding pockets. MHC class I molecules usually bind peptides around 9 residues, whereas class II molecules bind peptides of length 10-30 [4, 11]. It has been indicated that specific positions in this binding core, called anchor residues, are important for binding specificity and affinity. In the 9 residue binding core, the positions 1, 4, 6, 7 and 9 are the important ones [6]. There is a study showing that not only binding cores but also flanking residues towards the N and C-terminal of the peptide affect the binding stability and affinity [7]. Predicting the affinity of these peptides is important for vaccine design. It is also vital for understanding the roles of immune system in various diseases. Due to the variability of the locations of the class II peptide binding cores, predicting the affinity of these peptides is difficult. The task of the prediction algorithms has been learning the binding motif and using it for prediction. Various methods have been employed for this task. HMM, neural networks, Gibbs sampling, SVM and popular matrix based methods [1–3, 12, 15] are some of them. In this paper, we proposed a new method for predicting the affinity of the MHC Class II binding peptides based on their sequences. Our method classifies peptides as binding and non-binding. It is shown that some type of amino acids are preferred in some locations of binding peptides [6, 7], so we expect that some motifs are important for binding and they occur more frequently in binding peptide set and not in non-binding set. In order to find those frequent motifs we utilize n-grams and the information content of them. N-grams are subsequences of peptides composed of n consecutive amino-acids. They have recently been utilize for classification. Ganapathiraju and colleagues [5] investigated the n-grams in different organisms and observed that some n-grams are specific to some of the organisms. In addition, Vries et al. [9] utilize them to predict protein families. They have found most representative n-grams for each family and used that information for classifying proteins in a Bayesian probabilistic model. There are also reports that n-grams are successfully incorporated in GCPR ligand determination [16].

Our prediction method is based on a 3-step algorithm. In the first step we identify the informative n-grams, where $n \in [1, 5]$. We declare an n-gram as informative according to the frequency of the n-gram in the distributions of the both classes. In the next

step, the alphabet size is reduced such that each resulting amino-acid grouping to capture informative sub-groups. At the last step, by utilizing the informative n-grams, we aim to predict the class of a given sequence. In order to do this we employ two different prediction schemes. We have tested our methods on the MHC Bench IV-b data set [13]. Various other methods have been applied to this data set, and our methods perform better than most of these methods.

2. DECIDING INFORMATION CONTENT OF AN N-GRAM

Information content of an n-gram is decided after two steps: determination of classes, and, for each class finding the distribution of the n-grams. In the first step, the sequences having affinity less than or equal to 0 is assigned to the first class, *unbinding peptides*, and the rest of the proteins are assigned to the second class, *binding peptides*. In the second step for the given n, we find the distribution of n-grams for both classes.

We declare an n-gram as informative according to the cumulative distribution of the frequency of the n- gram lies into in the distributions of the both classes. The cumulative distribution function(*cdf*) of a real valued random variable X is defined as:

$$F(x) = P(X <= x)$$

For a specific n-gram, whose information content is explored, *cdf* is calculated for distributions of both classes. The *cdf* refers to the ratio of number of n-grams having frequency less than the explored n-gram. As an example, for n = 4, assume the 4-grams {AGIR, AGLH, KWVF, NCPA, and, DETY} show up in the sample with the following frequencies.

	AGIR	AGLH	KWVF	NCPA	DETY
Class-I	10	30	20	40	50
Class-II	10	40	50	30	20

So, the corresponding *cdf* of these 5 4-grams for both classes are as following.

	AGIR	AGLH	KWVF	NCPA	DETY
Class-I	0.2	0.6	0.4	0.8	1
Class-II	0.2	0.8	1	0.6	0.4

Then, the target space of the *cdf*,([0,1]), is divided into 3 subspaces according to the *minimum* and *maximum* thresholds as shown in Figure 1. The n-grams have a *cdf* less than the minimum threshold is in region 1 and the ones having a *cdf* between minimum and maximum thresholds are in the 2^{nd} region and the rest are in the 3^{rd} region. If the given n-gram is in the same region for both distributions then it is accepted as uninformative, otherwise informative. Assume that the minimum threshold is 0.25 and the maximum one is 0.75. For the above example, the 4-grams AGIR and AGLH are uninformative. For the informative n-grams, the absolute value of the *cdf*s from both classes is assigned as the information content of the n-gram. The 4-grams KWVF and DETY have the same information content whereas NCPA has a smaller value.

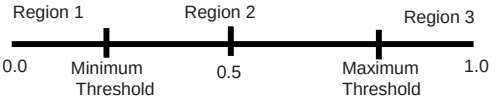

Fig. 1. Dividing the frequency space into 3 subspace according to the minimum and the maximum quartiles

3. CLUSTERING INFREQUENT AMINOACIDS: DECREASING THE ALPHABET SIZE

Since there is some preference on some positions of the binding peptides, it is expected that there will be a bias towards some types of amino acids in binding peptides. In a set of binding peptides this bias may not be apparent if we just look at the single aminoacid frequencies. However, if amino acids can be grouped based on their chemical and physical properties, reducing the size of the alphabet, it is possible to see that some of these abstract groupings become more frequent for a class. In order to find the sub-groups of amino acids, uninformative 1-grams are extracted, and among those uninformative 1-grams, the ones that are infrequent in binding and non-binding sets, are extracted. Extracted aminoacids grouped together if they are in the same group for Table 1. After grouping we recalculate frequencies with reduced alphabet, and repeat the procedure until there is no infrequent 1-gram.

Hence, the resulting groupings depend on the dataset and the ones given in Table 1. Assume that the aminoacids R and H are infrequent for both classes in the sample set, but, K has a *cdf* more than minimum threshold. Then, R & H are grouped in together where as, K will not join this group and be

analyzed separately.

Table 1. Possible Groupings

Group 1	D E
Group 2	R H K
Group 3	N Q S T
Group 4	F W Y
Group 5	C A P G V
Group 6	I L M

4. ALGORITHM

In the previous sections we described how to find the information content of an n-gram using the distribution of both classes and how to reduce the size of the alphabet. In this section, we describe the algorithms that utilize the informative subsequences to find the affinity class of a given aminoacid sequence. After identifying the employed n-grams, the cdf of these n-grams are summed up for each class and the one with the maximum sum is assigned as the predicted class for the given sequence.

4.1. N-gram Algorithm

Any subsequence of length n is taken into account. The information held by the informative length n subsequences is combined and the majority class is assigned as the class for the query sequence. For example, for n = 3, and the query sequence ACDEFVWYZ of length 9, there are 7 3-grams. The information content of these 3-grams are combined and assigned as the class of the query sequence, ACDEFVWYZ.

This approach can be utilize in two different ways. The first style is using only the n-grams for a fixed n, as shown in the above example. The other one is employing all the n-grams for which information content has been explored, for $n \in [1, 5]$. We named the second one as UAL(Utilize All n-grams).

4.2. Dynamic Approach

The algorithm shown in Table 2 is proposed for a problem which is a variant of the matrix multiplication problem(MMP). The MMP problem can be summarized as finding the order of multiplication such that the total cost of the multiplication operation between the matrices will be minimized and each matrix will be multiplied only once. The given algorithm utilizes the information content of all n-grams where $n \in [1, 5]$. For a given sequence the algorithm aims to find the division with the greatest information content. Hence, the objective is maximizing the sum of the information contents while dividing the given sequence into subsequences. For a sequence of ACD the algorithm considers the following subsets {ACD, A-CD,AC-D,A-C-D}. Clearly, as inherited from the matrix multiplication problem, this algorithm considers each aminoacid once. The only difference is that this solution also considers the size of the n-gram. The length multiplier, LM, is added to the algorithm for this purpose. The LM function takes the size of the n-gram as input and returns $1 + n/10$. Due to the space restrictions, we do not mention the elements of the dynamic programming.

Table 2. Dynamic Algorithm Using each Aminoacid Once

```
for i := 1 to n do
  Inf_{i,i} := LM(1)*Information Content(IC) of aminoacid_i

length = length of the given sequence
for n := 2 to 5 do
  for i := 1 to length-n do
    j := i + n
    Inf_All_{i,j}= LM(n)*IC of the n-gram_{i,j}
    Inf_Divide_{i,j} = Max_{i=<k<j}{Inf_{i,k} + Inf_{k+1,j}}
    Inf_{i,j} = Max{Inf_All,Inf_Divide}
```

5. RESULTS

There are 10 datasets available at [13]. An ideal test set should contain equal number of binders and non-binders. In absence of this, the evaluation parameters will show bias. Hence, we test our methods on the dataset 4-b, which is categorized under the label of balanced binders and non-binders by the owners. Raghava and Singh [13] evaluate 12 different approaches under 3 categories, motif, matrix and ann, on this dataset. We will give the rankings of our results among the given ones. Elimination process of the infrequent 1-grams has two iterations. At the end of the first iteration, the size of the alphabet is decreased from 20 to 18 and at the end of the last iteration it is reduced to 14. Experiments were done for each alphabet. The ordinary n-gram algorithm is run for $n <= 3$. The data set is divided into 5 equal partitions. Each partition is used as the test set, whereas the remaining are used as the training

set.

Table 3. Results for 4-b dataset

Alphabet Size → Methods↓	20	18	14
1-gram	0.546	0.6284	0.61
2-gram	0.5788	0.584	0.6045
3-gram	0.6627	0.6746	0.6954
UAL	0.6507	0.6832	0.6695
Dynamic	0.6421	0.673	0.6575

As shown in Table 3, each method performs better with the reduced size alphabets than the original alphabet. As the size of the alphabet decreases, the accuracy level for 2-gram and 3-gram algorithms increases. This is not the case for the others. They perform better with the alphabet of size 18. We pick the best 4 results from [13] and compare our results with the top performing methods. The selected results and the authors of the corresponding papers are depicted below.

Authors	Rammennsee et al. [14]	Marshal et al. [10]	Struniolo et al. [17]	Hammer et al. [8]
Accuracy	0.7003	0.6849	0.6764	0.6627

The highest level of accuracy achieved by our methods is 0.6954. Only one of the shown results, Rammennsee et al. [14], in [13] is greater than this. The highest accuracy achieved by Dynamic approach is 0.673 and this is slightly less than the third one, Struniolo et al. [17]. That of UAL method is 0.6832 and this is close to the second one, Marshal et al. [10].

6. CONCLUSION & DISCUSSION

Our methods have surpassed many other methods whose results are shown in [13]. We have shown that our genuine and simple approach is as accurate as those complicated methods.

We are currently investigating new algorithms and trying our existing algorithm on the other data sets. One possible modification to our dynamic algorithm might be the usage of every amino acid more than once. Another one to the same algorithm might be changing the LM function according to the performance of n-gram algorithm on various n. For the used dataset, 4-b, 3-gram performs the best, so the LM function can be designed such that the 3-grams are favored.

References

1. M. Bhasin and GPS. Raghava. Svm based method for predicting hla-drb1*0401 binding peptides in an antigen sequence. *Bioinformatics*, 20:421–423, 2004.
2. Vladimir Brusic, George Rudy, and Leonard C. Harrison. Prediction of mhc binding peptides using artificial neural networks. *In Complex Systems: Mechanism of Adaptation*, pages 253–260, 1994.
3. S. Buus, SL. Lauemoller, P. Worning, C. Kesmir, T. Frimurer, S. Corbet, A. Fomsgaard, H. Hilden, A. Holm, and S. Brunak. Sensitive quantitative predictions of peptide-mhc binding by a 'query by committee' artificial neural network approach. *Tissue Antigens*, 62:378–384, 2003.
4. Flora Castellino et al. Antigen presentation by mhc class ii molecules: invariant chain function, protein trafficking, and the molecular basis of diverse determinant capture. *Hum Immunol*, 54(2):159–169, 1997.
5. M. Ganapathiraju, J. Klein-Seetharaman, R. Rosenfeld, J. Carbonell, and R. Reddy. Comparative n-gram analysis of whole-genome protein sequences. In *Proceedings of the Human Language Technologies Conference*, 2002.
6. AJ. Godkin, T. Friede, M. Davenport, S. Stevanovic, A. Willis, J. Jewell, A. Hill, and H.-G. Rammensee. Use of eluted peptide sequence data to identify the binding characteristics of peptides to the insulin-dependant diabetes susceptibility allele hla-dq8 (dq 3.2). *Int. Immunology*, 9:905, 1997.
7. AJ. Godkin, KJ. Smith, A. Willis, MV. Tejada-Simon, J. Zhang, T. Elliott, and AVS. Hill. Naturally processed hla class ii peptides reveal highly conserved immunogenic flanking region sequence preferences that reflect antigen processing rather than peptide-mhc interactions. *Journal of Immunology*, 166:6720–6727, 2001.
8. Bono E. Gallazi F. Belunis C. Nagy Z. Hammer, J. and F. Singaglia. Precise prediction of major histocompatibility complex class ii peptide interaction based on side chain scanning. *J.Exp.Med.*, 180:2353, 1994.
9. JK. Vries JK, R. Munshi, D. Tobi, J. Klein-Seetharaman, PV. Benos, and I. Bahar. A sequence alignment-independent method for protein classification. *Appl Bioinformatics*, 3(2-3):137–48, 2004.
10. K.W. Marshal, K.J. Wilson, J. Liang, A. Woods, D. Zaller, and J.B. Rothbard. Prediction of peptide affinity to hla-drb1*0401. *J. Immunol.*, 154:5927–5933, 1995.
11. H. Max, T. Halder, H. Kropshofer, M. Kalbus, CA Muller, and H. Kalbacher. Characterization of peptides bound to extracellular and intracellular hla-dr1 molecules. *Hum Immunol*, 38:193–200, 1993.
12. M. Nielsen, C. Lundegaard, P. Worning, SL. Lauemoller, K. Lamberth, S. Buus, S. Brunak, and P. Lund. Reliable prediction of t-cell epitopes using neural networks with novel sequence representations. *Protein Science*, 12:1007–1017, 2003.
13. G.P.S. Raghava and Harpreet Singh. Evaluation og mhc binding peptide prediction methods. In *http://www.imtech.res.in/raghava/mhcbench*.
14. H. G. Rammensee, T. Friede, and S. Stevanovics. Mhc ligands and peptide motifs: first listing. *Immunogenetics*, 41:178–228, 1995.
15. HG. Rammensee, J. Bachmann, NPN. Emmerich, OA. Bachor, and S. Stevanovi. Syfpeithi: database for mhc ligands and peptide motifs. *Immunogenetics*, 50:213–219, 1999.
16. U. Sezerman, A. Akalin, Z. Kasap, and E. Kavak. Gpcr ligand determination using svms. In *ISMB/ECCB*, 2004.
17. T. Sturniolo, E. Bono, J. Ding, L. Raddrizzani, O. Tuercei, U. Sahin, M. Braxenthaler, F. Gallazzi, M.P. Protti, F. Sinigaglia, and J. Hammer. Generation of tissue-specific and promiscuous hla ligand database using dna microarrays and virtual hla class ii matrices. *Nat. Biotech.*, 17:555–561, 1999.