Prediction of MHC Class I Binding Peptides by a Query Learning Algorithm Based on Hidden Markov Models

K. UDAKA^{1,*}, H. MAMITSUKA², Y. NAKASEKO¹ and N. ABE^{2,3}

¹ *Department of Biophysics, Kyoto University*

² *Theory NEC Laboratory, RWCP (Real World Computing Partnership), c/o Internet Systems Research Laboratories, NEC corporation, Japan*

³ *IBM Thomas J. Watson Research Center, RO. Box 218 Yorktown, NY 10598*

(*author for correspondence; e-mail: keiko@molbio.biophys.kyoto-u.ac.jp)

Abstract. A query learning algorithm based on hidden Markov models (HMMs) is developed to design experiments for string analysis and prediction of MHC class I binding peptides. Query learning is introduced to aim at reducing the number of peptide binding data for training of HMMs. A multiple number of HMMs, which will collectively serve as a committee, are trained with binding data and used for prediction in real-number values. The universe of peptides is randomly sampled and subjected to judgement by the HMMs. Peptides whose prediction is least consistent among committee HMMs are tested by experiment. By iterating the feedback cycle of computational analysis and experiment the most wanted information is effectively extracted. After 7 rounds of active learning with 181 peptides in all, predictive performance of the algorithm surpassed the so far best performing matrix based prediction. Moreover, by combining the both methods binder peptides ($\log K_d < -6$) could be predicted with 84% accuracy. Parameter distribution of the HMMs that can be inspected visually after training further offers a glimpse of dynamic specificity of the MHC molecules.

Key words: algorithm, binding, experimental design, MHC class I molecules, peptides, prediction, query learning, specificity, string analysis

1. Introduction

Major Histocompatibility Complex (MHC) class I molecules present peptides to cytotoxic T lymphocytes (CTLs) for immune surveillance. A given MHC class I molecule binds a limited but relaxed specificity of peptides. There exist approximately one or two peptides in every 200 random peptides. The hallmark of the MHC class I and class II genes is their extensive polymorphism. Allelic variations are mostly concentrated in the peptide binding cleft on top of the MHC molecule and they have often serious impacts on peptide selection. In order to understand the nature of MHC bound peptides and design effective T cell vaccines it is pivotal to understand the specificity of the given MHC molecule. Since the elucidation of major anchor amino acids by pool sequencing [1, 2] efforts have been focused on

unraveling more detailed specificity of the MHC molecules. They include identification of secondary anchors [3], side chain scanning [4, 5], positional scanning with peptide libraries [6–8] and computational analyses [9–11]. Based on these findings two sets of predictive algorithms, BIMAS [4] and SYFPEITHI [12] have been developed. However, information gained by the pool sequencing is position independent and it is hard to translate into quantitative terms as revealed by the limited success in prediction [13]. Previously, we have developed a positional scanning method with peptide libraries. Library scanning yields detailed profiles of amino acid preferences at every position of the peptide in quantitative values. This has allowed an automated prediction [14]. However, it is also position independent and predictions can be off when amino acids behave interactively, e.g. by inducing conformational adaptations on peptides or the MHC molecules [8, 14, 15].

Computational approaches have been introduced to extract not only the position specific information, but also some string patterns on peptides [16, 17]. However, to train an algorithm a large number of experimental data are needed. To date, the source of MHC binding peptides has been the databases deposited from the literatures where peptide bindings were determined by different methods and criteria [8, 12]. Therefore, the use of data has been limited to yes/no binaries on peptide binding. This, together with the limited number of data has been hampering the full exploitation of databases.

One of the authors has previously developed a supervised learning algorithm based on hidden Markov models and analyzed the HLA A*0201 binding peptides from MHCPEP [18] in binaries [16]. The effectiveness of the approach has prompted us to extend the method to predict the degree of peptide binding in real-number-values. However, training the algorithm inevitably requires a large number of binding data. For 9-mer peptides that are most common MHC class I ligands in human and mouse the sequence variation reaches $20^9 \approx 5 \times 10^{11}$. One or two peptides in every 200 random peptides bind with high affinity to a given MHC class I molecule [7, 8]. Therefore, it is pivotal to develop a method for extracting information on MHC binding properties of peptides from a small number of experiments.

Here, we propose a novel method for dynamic design of experiments based on a query learning to realize data- and computationally-efficient learning. It is an iterative feedback approach combining a computational analysis and experiments.

2. Results

2.1. A DYNAMIC DESIGN OF EXPERIMENTS BY A QUERY LEARNING ALGORITHM WITH HIDDEN MARKOV MODELS

We introduce a query learning algorithm called ‘Query by Bagging (Qbag)’ whose performance has been demonstrated superior in several applications [19, 20]. Qbag is a committee-based query learning method combined with arbitrary supervised learning algorithms as its components. It was chosen for the following advantages:

Figure 1. Hidden Markov Models trained by the peptide binding data. **a.** Basic structure of HMM designed in this study. Each state, represented by an open circle designates an amino acid position. The state bears symbol generation probabilities for every amino acid, i.e. the probability that an amino acid frequents that state. In order to tolerate lateral shifts of sub-strings by an amino acid, one position was defined as a variable state. This position can be occupied by the first or last amino acid. The core states of 8 amino acids are connected by one-directional transitions. The first and 8th states have alternative state transition probabilities connected to initial/final state or to the variable state. Six independent paths were designed without inter-connection. This simple model exploits the fact that, for MHC class I binding peptides, orientation is fixed and size variation is limited and it saves the data needed for training. When an HMM is trained with a subset of the peptide binding data, symbol generation probabilities and state transition probabilities are optimized so that the log-likelihood values of given peptides approximate their log K_a values. **b. c.** Two examples of the trained HMMs. Three most frequented paths are presented. Only the symbol generation probabilities and state transition probabilities that exceed 0.1 (in bold for > 0.3) are shown.

Procedure: Query-by-Bagging (Qbag)

Input Parameters: M, A, T, R, I, D

Obtain initial sample $S_1 = \langle (x_1, y_1), \dots, (x_I, y_I) \rangle$.

For $i = 1, \dots, M$

1. By re-sampling from S_i with uniform distribution, obtain sub-samples $S'_1, \dots, S'_T$ (of same size as S_i).
2. Run component learning algorithm A on $S'_1, \dots, S'_T$ to obtain hypotheses $h_1, \dots, h_T$.
3. Randomly generate R examples (peptides) C_i .
4. For all examples (peptides) x in C_i , compute the variance $v(x)$ by:

$$v(x) = \frac{\sum_{t=1, \dots, T} |h_t(x) - h_{ave}(x)|^2}{T}$$
where $h_{ave}(x) = \frac{\sum_{t=1, \dots, T} h_t(x)}{T}$
5. Select D examples (peptides) $E_i = \{x_1^*, \dots, x_D^*\}$ from C_i having the largest values of $v(x)$.
6. Empirically determine the binding values of $E_i, y_1^*, \dots, y_D^*$ and update the training data by
 $S_{i+1} = \text{append}(S_i, \langle (x_1^*, y_1^*), \dots, (x_D^*, y_D^*) \rangle)$

End For

Output: Obtain final hypothesis by taking average of the T hypotheses $h_1, \dots, h_T$ from the last (M -th) iteration.

Figure 2. Outline of the dynamic design of experiments by Qbag.

1. Data can be labeled with real numbers. 2. It places no restriction on the type of learning algorithm as its component. For the component learning algorithm, we chose a supervised learning algorithm for hidden Markov models (HMMs). HMMs have been applied to multiple alignment, since it can capture a string pattern common to the given multiple sequences even if it appears at different positions [21, 22]. Applications of HMMs extend to prediction of protein structures [23] and ORF finding in the genome [24]. HMM was chosen for the following reasons. 1. It can visualize amino acid preferences and sequence patterns hidden in a given set of training peptides. This is not possible with artificial neural network (ANN). 2. Learning and predictive abilities are at least comparable with other learning models such as ANN. Since general unsupervised learning algorithms for HMMs could only deal with data having only one label we employ a supervised learning algorithm due to Mamitsuka [16]. A basic structure of HMM designed is shown in Figure 1a.

Progress of learning by the Qbag algorithm with HMMs.

Figure 3. Progress of Qbag learning. Experimentally measured binding abilities of peptides are shown. The data are displayed in relation to the previously described library score because it helps to estimate the binding abilities. a. Initial binding data of 42 peptides that include known or actively sought binders as well as unrelated peptides. b. A part of the random peptides used for the first round learning by Qbag. c. Peptides selected for experimental verification after the first round of Qbag learning. d. Binding abilities of peptides that are selected for binding measurement after the 7th round of learning.

2.2. LEARNING PROCESS

The Qbag designed experiments proceeded through the following steps (Figure 2).

Step 1 and 2: Training of HMMs with subsets of the binding data:

D^b binding ability of 148 peptides of 9-mer was measured (Figure 3a, b). Initial data of 42 peptides included known binders, unrelated peptides or those that were actively sought using library scanning [14] (Figure 3a). Binding data of 106 random peptides were added in 4 experiments (Figure 3b). The initial learning was conducted using these 148 binding data. To train a multiple number of HMMs which will collectively serve as a committee for prediction, 148 peptides were sampled at random with replacement from the data pool and one of the fifty HMMs was trained with an independently sampled subset of the data. This process was repeated 50 times. As a result, fifty independently trained HMMs were obtained, each with slightly different parameter settings. From the second round learning the data newly analyzed in Step 6 are added to the data pool for Step 1.

Step 3: Generation of test peptides.

In order to search the universe of 20^9 for D^b binding peptides, 100 000 9-mer peptides were created randomly at a time.

Step 4,5: Judgement by the committee HMMs

Fifty independently trained HMMs were given a task of predicting the binding ability for 100 000 random peptides. One peptide whose predictions were least consistent among committee HMMs, i.e. bearing the largest variance, was selected. By repeating Steps 3–5 for 50 times 50 such peptides were selected. In effect, ~ 5 million peptides (1/100 000 of the whole universe) could be screened.

Step 6: Measurement of D^b binding and feedback to the database

24–36 peptides were arbitrarily chosen from the 50 selected peptides and subjected to measurement of D^b binding abilities. The results were deposited to the data pool used in Step 1.

As shown in Figure 3c, the first round learning dramatically changed the profile of peptides. With random peptides binders could be rarely found that were 10^{-6} (M) or below in Kd (Figure 3b). However, after the first iteration of learning, most of the selected peptides bound more or less to D^b . The fact that peptides bearing a wide range of binding abilities are equally selected has a crucial impact on information gathering. Peptides in the marginal zone can often give fine details because even a smallest change can have a serious impact on binding. By excavating the most unpredictable peptides with random sampling the most wanted information can be sought efficiently. In effect, it preferentially explores unknown rules in the unexamined points in the data space. We repeated the learning cycle for 7 rounds. In all 181 peptides were synthesized and tested. The profile of the 7th round (Figure 3d) appears not much different from the first round (Figure 3c), i.e. harboring peptides ranging from non-binders to the highest binders. This indicates that an ample room is still left for further improvement by iteration.

2.3. ASSESSMENT OF THE PREDICTIVE POWER OF TRAINED HMMS

In order to assess the effectiveness of the Qbag learning method, a program for prediction was constructed using 329 binding data obtained above (148 before learning, 181 after learning). First, one hundred HMMs loaded with different initial parameters were trained independently with subsets of 329 binding data as described in Step 1,2. Two of the trained HMMs are shown (Figure 1b, c). Test peptide was scored by HMMs and given the average by 100 predictions.

We compared the predictive capability of the HMMs with a so far best performing matrix method [14]. The latter is calculated from the positional scanning data

Figure 4. Prediction of binding abilities for random peptides. Forty-one peptides each were randomly selected from the peptides scored within the highest 1% by the Qbag or Library method. Peptide binding was measured by stabilization of D^b. Peptides dually scored high by the two methods (●) and those scored high only by one method (○) are shown. a. Peptides scored high by the Qbag method. b. peptides scored high by the library method. c. Precision-Recall (P-R) curves for Lib high scored peptides. Solid line designates precision, i.e. the fraction of peptides that are proven to be binders up to that peptide arranged in the descending order by the Lib score (log Kd < -6). Broken line indicates recall, i.e. the fraction of actual binders among 41 peptides tested. d. P-R curves for Qbag ranked peptides.

analyzed by the combinatorial peptide libraries and is referred to as the Library method.

2.4. PREDICTIVE POWER OF QBAG ON RANDOM PEPTIDES

In order to make fair judgement on unbiased population we next analyzed the predictive power on random peptides. One hundred HMMs trained as above were used to predict one million peptides. Forty-one peptides each were randomly sampled from the top 1% (10 000) scored peptides by the Qbag or Lib method. Nineteen Que^{hi} and 19 Lib^{hi} peptides were ranked within 1% by the both methods. As shown in Figure 4a, b and Table I, Qbag exhibits higher binding. Precision-Recall curves in Figure 4c, d further indicate better linearity and coverage. The Library method tends to overestimate the binding for some peptides. Although Qbag alone can realize the unprecedented 61% (25/41) precision considering the Lib score effectively eliminates false positives and that improves the precision to 84% (16/19) with

Table I. Prediction of D^b binding peptides for random peptides

peptides	mean log Kd (M)	number of peptides with log Kd < −6	
<hr/>			
Que ^{high} sampled peptides			(%)
total	−6.08	25 / 41	61
Que ^{high} & Lib ^{high}	−6.59	16 / 19	84
Que ^{high} but Lib ^{low}	−5.64	9 / 22	41
<hr/>			
Lib ^{high} sampled peptides			
total	−5.78	16 / 41	39
Lib ^{high} & Que ^{high}	−6.03	9 / 19	47
Lib ^{high} but Que ^{low}	−5.56	7 / 22	32

Forty-one peptides ranked within 1% by the Qbag or Library method were examined for their MHC binding abilities. No identical peptide was sampled twice by the both methods.

the expense of binder coverage to 70% (16/25). This time, training of HMMs was halted at 181 peptides but the prediction would improve further if the training were continued. This is an impressive performance for a string analysis. Kearns and Vazirani have estimated that at least $\geq 1\,000$ data are necessary for a matrix based PAC learning in binaries with an error rate of 15% in the data source and confidence of 95% [9, 25]. 350–500 binding data were needed that were highly enriched with binder peptides to derive a reasonable matrix in binaries [9]. 300–400 data highly enriched with binders were also needed for ANN analysis by Gulukota *et al.* [10].

3. Discussion

Dynamic experimental design by Qbag is shown to be quite effective in string analysis of MHC binding peptides and it has surpassed the so far best performing matrix-based Library method in precision and coverage. This is also only the second program after the Library method [14] that can predict the binding capacities in real number values. For most investigators looking for MHC binding peptides it is not easy to measure the peptide binding. However, with this level of precision, direct screening of responsive T cells may become realistic. Once the HMMs are trained prediction can be automated for high throughout analyses. Even a genome wide analysis is possible. If this were extended to other MHC molecules, especially to human MHC (HLA) molecules, it would enable finding candidate peptides for immunotherapy against tumors and pathogens such as HIV and hepatitis C virus. In principle, it is possible to apply Qbag to MHC class II binding peptides. If so, a genome wide search for peptide antigens may prove effective in controlling autoimmune diseases.

So far a major obstacle for string analysis such as ANN, has been the insufficient number of binding data. Reaching this level of performance with 181 peptides by the introduction of Qbag seems promising. Unlike ANN, HMM has the advantage that the probability distribution and its string patterns in the hidden layer can be inspected visually after the training. As seen in Figure 1b, c, the patterns are in fact, quite interesting. D^b has the major anchor amino acids of N at position 5 (P5) and M or I at P9 [1]. Preferences for these amino acids can be seen more or less in the major paths. We tentatively designed the HMM with six paths. There turned out to be two or three major paths which peptides would take frequently. So restricting to six paths in the model effectively buys computational efficiency at no significant loss in representational power. The probability distribution in the HMM itself represents a dynamic specificity of the MHC molecule. It would be interesting to use the information to aid computer modeling in the future.

It was found that the Qbag and Library methods are complementary to each other and that combining them enhances predictive performance of either method in isolation. It appears that with the limited number of binding data, the trained HMMs have not yet learned as much position specific information as the Library method. This indicates that the method of randomizing the amino acid positions except for the interested position is in fact, quite effective for extracting position specific information [6, 7]. Library scanning also requires a set of 172 sub-libraries ($19 \times 9 + 1$, for a random reference library). However, the same set can be used to analyze any 9-mer binding MHC class I molecules. Therefore, for analysis of a new MHC molecule, combining the two methods with a couple of hundred new peptides for the Qbag learning would be a good option for parsimony.

There are possibilities for further improvement of Qbag experimentation: 1. Further iteration. 2. Modifying the HMM model. 3. Increasing the number of the final HMMs used for prediction. 4. Replacing HMM as the component learning algorithm by other algorithms. It is an advantage of Qbag that it is a versatile algorithm that can take an arbitrary learning algorithm as the component algorithm. Applying Qbag learning to other multi-parameter problems, such as drug design and search of cognate ligands would also be interesting future research topics.

4. Experimental Protocol

4.1. SUPERVISED LEARNING ALGORITHM FOR HMM

General scheme of the algorithm has been described [26]. Experimentally determined log K_d values were related to the HMM parameters as follows. Let H be an HMM and O be a training peptide. Let L_{OIH} be the log-likelihood of O for a given H. Set a constant, C as follows: $C = L_{OIH'} - L_{Ka}$, where L_{Ka}, is the average log Ka and H' is the reference HMM having uniform probability distributions.

1. Training HMM

Let L_{OIH}^* be the target log-likelihood of O in a given H. It is calculated as follows: $L_{\text{OIH}}^* = C + L_{\text{Ka}}^{\text{O}}$, where L_{Ka}^{O} is log Ka of the peptide O.

A function g_{OIH} is defined as follows:

$$g_{\text{OIH}} = (D_{\text{max}} - D_{\text{OIH}})/D_{\text{max}},$$

where $D_{\text{OIH}} = |L_{\text{OIH}}^*|^2$. D_{max} is a constant which satisfies $D_{\text{max}} > \max_{\text{O}} D_{\text{OIH}}$.

Probability parameters of H are trained so as to minimize the energy function, E that is based on a gradient descent algorithm.

$$E = \Sigma_{\text{O}} - \log g_{\text{OIH}}.$$

2. Prediction of log Ka for a peptide

The log Ka of a peptide, O is estimated by $L_{\text{OIH}} - C$.

4.2. PEPTIDES

Peptides were synthesized by Fmoc chemistry and purified by HPLC to the purity of > 95%. Peptides were analyzed by MALDI-TOF mass spectrometer (Voyager DE-RP, PerSeptive, MA). Concentration of peptide was determined by MicroBCA assay (Pierce, Rockford, IL) using BSA as standard. For rare peptides that react poorly with the BCA reagent a quantitative ninhydrin reaction was used, instead.

4.3. PEPTIDE BINDING

Peptide binding was measured by stabilization assay using RMA cells [8]. Brief 1×10^5 RMA cells incubated overnight at 260 °C were mixed with grading concentrations of peptide in 0.25% BSA containing DMEM, incubated 30 min at room temperature and then exposed to 37 °C for 70 min. The remaining D^b molecules on the cell surface were stained with FITC labeled anti-D^b mAb, B22.249 and analyzed by FACScan (Becton-Dickinson Japan, Tokyo). Differences between experiments were normalized by always including the D^b binding reference peptides [14].

Acknowledgements

We thank Mr R. Yamamoto for synthesis of a part of the peptides. This work was supported by RWCP sponsored by the ministry of economy, trade and industry. Support from Hayashi Memorial Foundation for Female Natural Scientists is also acknowledged.

References

1. Falk, K., Roetzschke, O., Stevanovic, S., Jung, G. and Rammensee, H.: Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules, *Nature* **351** (1991), 290–296.

2. Rammensee, H., Friede, T. and Stevanovic, S.: MHC ligands and peptide motifs: first listing, *Immunogenetics* **41** (1995), 178–228.
3. Ruppert, J., *et al.*: Prominent role of secondary anchor residues in peptide binding to HLA-A2.1 molecules, *Cell* **74** (1993), 929–937.
4. Parker, K., Bednarek, M. and Je, C.: Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains, *J. Immunol.* **152** (1994), 163–175.
5. Hammer, J., *et al.*: Precise prediction of MHC class II-peptide interaction based on peptide side chain scanning, *J. Exp. Med.* **180** (1994), 2353–2358.
6. Stryhn, A., *et al.*: Peptide binding specificity of major histocompatibility complex class I resolved into an array of apparently independent subspecificities: quantitation by peptide libraries and improved prediction of binding, *Eur. J. Immunol.* **26** (1996), 1911–1918.
7. Udaka, K., Wiesmuller, K.-H., Kienle, S., Jung, G. and Walden, P.: Tolerance to amino acid variations in peptides binding to the MHC class I protein H-2Kb, *J. Biol. Chem.* **270** (1995), 24130–24134.
8. Udaka, K., Wiesmuller, K.-H., Kienle, S., Jung, G. and Walden, P.: Decrypting the structure of MHC-I restricted CTL epitopes with complex peptide libraries, *J. Exp. Med.* **181** (1995), 2097–2108.
9. Brusic, V., Schoenbach, C., Takiguchi, M., Ciesielski, V. and Harrison, L.: Application of genetic search in derivation of matrix models of peptide binding to MHC molecules, *ISMB* (1997), 75–83.
10. Gulukota, K., Sidney, J., Sette, A. and DeLisi, C.: Two complementary methods for predicting peptides binding major histocompatibility complex molecules, *J. Mol. Biol.* **267** (1997), 1258–1267.
11. Honeyman, M., Brusic, V., Stone, N. and Harrison, L.: Neural network-based prediction of candidate T-cell epitopes, *Nat. Biotechnol.* **16** (1998), 966–969.
12. Rammensee, H., Bachmann, J., Emmerich, N., Bachor, O. and Stevanovic, S.: SYFPEITHI: Database for MHC ligands and peptide motifs, *Immunogenetics* **50** (1999), 213–219.
13. Andersen, M., *et al.*: Poor correspondence between predicted and experimental binding of peptides to class I MHC molecules, *Tissue Antigens* **55** (2000), 519–531.
14. Udaka, K., *et al.*: An automated prediction of MHC class I-binding peptides based on positional scanning with peptide libraries, *Immunogenetics* **51** (2000), 816–828.
15. Madden, D., Garboczi, D. and Wiley, D.: The antigenic identity of peptide-MHC complexes: A comparison of the conformations of five viral peptides by HLA-A2, *Cell* **75** (1993), 693–708.
16. Mamitsuka, H.: Predicting peptides that bind to MHC molecules using supervised learning of Hidden Markov Models, *Proteins* **33** (1998), 460–474.
17. Brusic, V., Rudy, G., Honeyman, M., Hammer, J. and Harrison, L.: Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network, *Bioinformatics* **14** (1998), 121–130.
18. Brusic, V., Rudy, G., Kyne, A. and Harrison, L.: MHCPEP – A database of MHC-binding peptides: update 1995, *Nucl. Acids Res.* **24** (1996), 242–244.
19. Abe, N. and Mamitsuka, H.: *In the fifteenth international conference on machine learning*, Morgan Kaufmann, Madison, Wisconsin, 1998.
20. Mamitsuka, H. and Abe, N.: *In the seventeenth international conference on machine learning*, Morgan Kaufmann, Stanford, Ca., 2000.
21. Baldi, P., Chauvin, Y., Hunkapiler, T. and McClure, M.A.: Hidden Markov models of biological primary sequence information, *Proc. Natl. Acad. Sci. USA* **91** (1994), 1059–1063.
22. Krogh, A., Brown, M., Mian, I.S., Sjolander, K. and Haussler, D.: Hidden Markov models in computational biology: applications to protein modeling, *J. Mol. Biol.* **235** (1994), 1501–1531.
23. Park, J., *et al.*: Sequence comparisons using multiple sequences detect twice as many remote homologues as pairwise methods, *J. Mol. Biol.* **284** (1998), 1201–1210.

24. Krogh, A., I.S., M. and D., H.: A hidden Markov model that finds genes in *E. coli* DNA, *Nucl. Acids Res.* **22** (1994), 4768–4778.
25. Kearns, M. and Vazirani, U.: *An introduction to computational learning theory*, MIT Press, Cambridge, Ma., 1994.
26. Mamitsuka, H.: A learning method of hidden Markov Models for sequence discrimination, *J. Comput. Biol.* **3** (1996), 361–363.