Available online at www.sciencedirect.com
SCIENCE DIRECT®
Enzyme and Microbial Technology xxx (2003) xxx–xxx
ENZYME and MICROBIAL TECHNOLOGY
www.elsevier.com/locate/enzmictec

Hidden Markov model-based approach as the first screening of binding peptides that interact with MHC class II molecules

Ryuji Kato, Hideki Noguchi[1], Hiroyuki Honda, Takeshi Kobayashi*

Department of Biotechnology, School of Engineering, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan

Received 11 February 2003; accepted 8 April 2003

Abstract

The immune system is initiated and regulated through a process starting from the binding of antigenic peptides to major histocompatibility complex (MHC) molecules. Detailed understanding of such interactions would lead to the development of vaccine design for infectious diseases, and immunotherapies for autoimmune diseases and cancer. Since MHC class II genes are highly polymorphic, a computational prediction tool for the first screening of antigenic peptides that bind to MHC class II molecules, is highly desirable. In the present study, hidden Markov model (HMM) was applied for the screening of peptides that interact with nine MHC class II molecules, specifically, human leukocyte antigen (HLA)-DR1, -DR2, -DR4, -DR7, -DR11, -DR15, -DR17, -DR51, and -DQ2. When high-binding peptides interacting with each MHC molecule were subjected to the constructed HMM model, significantly high likelihood values were obtained, as compared to the non-binding peptides as negative control. With the receiver-operating characteristic analysis for the prediction evaluation, our model showed high prediction accuracy, with an average AUC value of 0.87 for all molecules. The HMM model that was trained by HLA-DQ2 showed significantly low likelihood values to peptides that bind to eight HLA-DR molecules. This suggests the high potency of our HMM model for discriminating HLA-DQ binding peptides from HLA-DR binding peptides.

Keywords: MHC class II; Binding peptide; Binding prediction; Hidden Markov model; Bioinformatics

1. Introduction

Adaptive immune system is a self-defense mechanism against most microbial or viral invasions that have escaped the innate immune system. The system targets infectious invasion sources with proteins or peptides that originating in the antigens, and then accomplishes the defense using antibodies and killer cells [1,2].

Major histocompatibility complex (MHC) molecules are peptide receptors that play a central role in the initiation and regulation of T cell-mediated immune responses [1,2]. These immune processes involve the binding of peptides to MHC molecules inside the antigen presenting cells (such as B cells, macrophages, and dendric cells), the transport of these complexes to the cell surface, and the presentation for the recognition by T cells [2]. A peculiarity of MHC genes is the extensive polymorphism, characterized by the presence of hundreds of allelic variants. Each variant of the MHC molecules binds with foreign and self-antigenic peptides, in accordance with the specific binding motif, thus mediating the individual differences in immune responses [1]. Immune response is initiated if the strength of the interactions surpasses a certain threshold, regardless of the origin of the bound peptide (i.e. self-peptide or non-self-peptide). Malfunctions, such as allergy or severe autoimmune diseases, are also caused in a similar manner. In a positive implication, an effective presentation of cancer antigenic peptides should lead to immunological therapy [3]. Therefore, detailed understanding of the interaction mechanisms that govern the binding and selection of the peptides is essential for the development of vaccines against infectious diseases and immunotherapies for allergy, autoimmune diseases, and cancer [3,4].

MHC molecules are classified as MHC class I (HLA-A, -B, and -C) and class II (HLA-DR, -DQ, and -DP). Class I molecules accommodate peptides with a narrow distribution in their lengths (8–11 residues [5]), and the bound peptides

Abbreviations: MHC, major histocompatibility complex; HLA, human leukocyte antigen; HMM, hidden Markov model; SSS, successive state splitting; ROC, receiver-operating characteristic; AUC, area under the ROC curve

* Corresponding author. Tel.: +81-52-789-3213; fax: +81-52-789-3214.
E-mail address: takeshi@nubio.nagoya-u.ac.jp (T. Kobayashi).
[1] Present address: Genome Informatics Team, Human Genome Research Group, Genome Science Center, RIKEN, 1-7-22 Suehirocho, Tsurumi-ku, Yokohama 230-0045, Japan.

0141-0229/03/$ – see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/S0141-0229(03)00150-9

are recognized by cytotoxic T cells. The binding pockets are known to accommodate nine anchoring amino acids as the binding core-regions [6]. Based on such clear structural binding rules, peptides that can bind to a MHC class I molecule are easily identified by searching the binding core-region.

Class II molecules are encoded by genes in the subregions, such as DR, DQ, and DP [1,2]. For example, MHC molecules that are associated with the DR genes are designated as HLA-DR1. HLA-DQ molecules differ from HLA-DR molecules in several aspects, such as high polymorphism in both the polypeptide chains that comprise the MHC class II molecule, or its expression manner [7]. Class II molecules bind with longer peptides than their class I counterparts, as recognized by helper T cells. Since the specific hydrophobic residues that close the ends of the class I peptide binding groove are absent, the length of the bound peptides are unrestricted [8]. Accordingly, the peptides that bind to class II molecules vary from 11 to 30 residues [2]. Although binding pockets are also observed in MHC class II molecules [9], positional differences among the alleles impose diverse constraints on the peptide sequences that may bind to the class II molecule [10]. It was confirmed that peptides with nine residues, involving several anchor amino acids, bind to the class II molecule groove—however, their immunogenic strengths vary greatly by the addition of amino acids to either or both ends of the amino and carboxyl termini [11]. With such wide distributions in the lengths of the binding peptides, and differences in allele specific binding pocket positions or the residues outside the core-region, identification of MHC class II binding peptides have proven to be a monumental task.

Several predictions of the antigenic peptides that bind to several allelic MHC class II molecules have been attempted with characteristic structures in peptide [12,13], and with computational approaches including neural network [14], evolutionary algorithm, and artificial neural network [15], iterative stepwise discriminant analysis meta-algorithm [16], and fuzzy neural network [17]. Correlations were observed between the binding strengths and the characteristics of amino acids at certain positions, in which predictive scores for the peptides were calculated from the quantitative matrix showing preference of amino acids for each position in the MHC-binding peptides [12,13]. Although computational prediction approaches [14,15,17] have shown high prediction accuracy, time-consuming experiments, usually made only for few molecules, are still necessary in constructing a quantitative matrix.

Investigations of peptides that strongly bind to MHC class II molecules, which effectively regulate the immune response, may result in a peptide vaccine against infectious microorganisms, a cancer-targeting peptide for cancer immunotherapy, or a peptide drug for the autoimmune diseases. To this end, bioinformatic tools can provide the methods for the first screening in the exhaustive drug discovery of various peptides that can interact with MHC class II molecules. Since the variation of peptide sequences are enormous in an exhaustive research, effective but simple screening tools that require fewer experimental processes are highly desirable. As an example, even with short peptides with four residues, 20^4 (=160,000) variants should be involved in an exhaustive investigation for their activity.

Previously, we have demonstrated that the hidden Markov model (HMM) can be effectively adapted for the binding prediction of peptides that interact with MHC class II molecules, with the combination of successive state splitting (SSS) [18]. HMM is a stochastic model, which is suitable for representing time-series and biological sequences, and has been widely used in the field of bioinformatics [19]. An important feature of HMM is the flexibility that provides the representation of peptide sequences of various lengths within a single model, and accordingly, some binding sequence rules underlying the full length of binding peptides are expected to be conserved in the training of the model. In our previous work [18], a model (S-HMM) has showed high prediction accuracy in the binding peptide prediction of human MHC class II molecule, HLA-DR1. As a continuation of our studies of the extensive polymorphic MHC molecules, the universal predictive power of S-HMM should be established, in addition to HLA-DR1. Herein, we describe our investigations to establish the universal predictive power of S-HMM using the binding peptide data of the other eight MHC class II molecules (HLA-DR2, -DR4, -DR7, -DR11, -DR15, -DR17, -DR51, and -DQ2). The model structures are optimized, and the characteristics of acquired models are presented.

2. Materials and methods

2.1. Peptide data

Peptide amino acid sequence data and their binding strength to human MHC class II molecules were obtained from the MHCPEP database (http://wehih.wehi.edu.au/mhcpep/) [19]. From the database, 305 sequences (HLA-DR1), 33 sequences (HLA-DR2), 127 sequences (HLA-DR4), 34 sequences (HLA-DR7), 67 sequences (HLA-DR11), 46 sequences (HLA-DR15), 33 sequences (HLA-DR17), 39 sequences (HLA-DR51), and 46 sequences (HLA-DQ2) were selected as high-binding peptides. The peptide lengths in the resulting set ranged from 9- to 25-mer, and the peptide length distribution of all the allelic molecules is shown in Table 1.

2.2. Hidden Markov model

The left-to-right HMM used in this study has basically the same structure as described in our previous work [18]; however, for clarification, a brief explanation is given as follows.

Table 1
Peptide length distribution of high-binding peptides data to HLA allelic molecules

	Peptide length (number of residues)																	Total
	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	
DR1	120	0	25	4	57	5	59	11	5	4	4	8	0	0	0	1	2	305
DR2	0	0	0	0	0	1	17	3	0	1	2	7	2	0	0	0	0	33
DR4	60	0	0	0	54	1	8	0	0	0	0	2	0	0	0	2	0	127
DR7	0	0	0	0	4	4	16	1	1	2	1	3	0	0	1	1	0	34
DR11	53	2	0	0	0	7	4	1	0	0	0	0	0	0	0	0	0	67
DR15	0	0	4	3	3	3	16	3	2	3	4	3	0	0	0	1	1	46
DR17	0	0	0	0	1	2	13	9	3	2	1	0	0	0	0	2	0	33
DR51	0	0	0	0	30	1	1	0	0	1	0	5	0	0	0	1	0	39
DQ2	0	21	4	1	2	5	2	6	1	1	0	2	0	0	0	1	0	46

HMMs are stochastic models that can model processes, represented by sets of various sequence of symbols. The sequences of symbols typically represent protein, DNA, or RNA sequences in the bioinformatic field. HMM is known to encode properties of the sequences in various lengths within the same class. In the basic HMM [20], the model is represented by a finite set of nodes called "states," and arcs connecting the states. Every state is associated with outputs (in the present case, the amino acid character) to be generated according to the involving probability named symbol generating probabilty in this work. Every arc showing the transition from a state to the next state is governed by a set of probabilities called transition probabilities.

An example of left-to-right HMM used in the present study is shown in Fig. 1. The HMM consists of linear lined states connected with arcs in one direction. The starting and the finishing states do not emit any amino acid. A sequence of observations (a peptide sequence) and the probability of generating such sequence from the model can be calculated as follows: in Fig. 1, five peptides of various lengths, with partially similar amino acid sequences (TKH, TKQQQ, TKQHHQHQHQQ, TKHQQQQQQHHHQHH, and TKHHHHHHH HHHQHHHHHHHHHHQHQ), were modeled by one simple structure using HMM. All five peptides have identical amino acids in the first (T: threonine) and second (K: lysine) positions. Starting from the third position, the peptides consist of either H (histidine) or Q (glutamine), with different orders and repeated times. The amino acid distribution from the third position can be expressed by only one state, which is state 3 in the model. State 3 may represent 50 amino acids with the distribution of H and Q symbols (H = 32, Q = 18), and the symbol generation probabilities can be given as 32/50 = 0.64 (H) and 18/50 = 0.36 (Q). To express the length distribution of the sequences, the use of state 3, repetitively utilized for a total of 30 times, may allow the model to represent all the

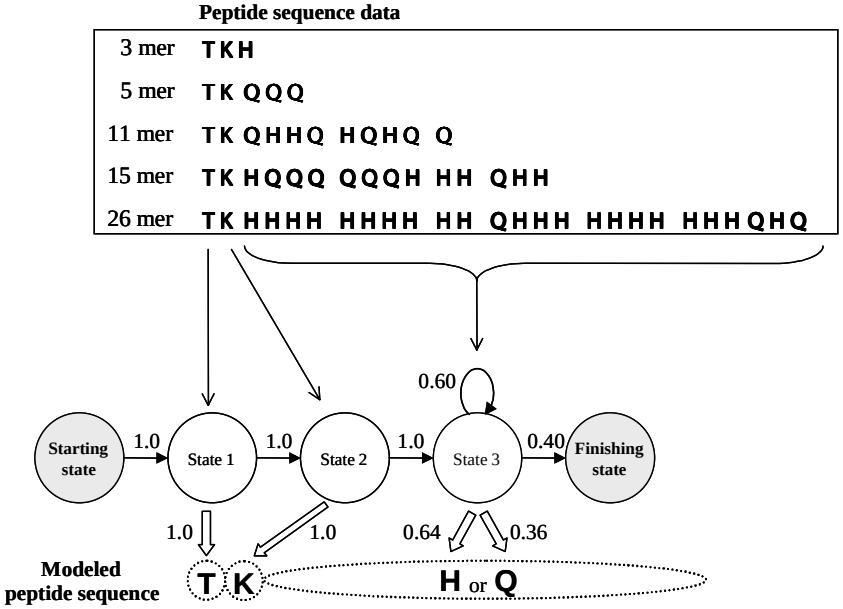

Fig. 1. An example of left-to-right HMM using the training data consisting of five peptide sequences with differing lengths. Symbols: T, threonine; K, lysine; H, histidine; Q, glutamine; normal arrows, transition probability; open arrows, symbol generation probability.

different sequence lengths. For example, the sequence TKQHHQHQHQQ uses the state 3 twice, repeatedly, from the 4th to the 5th amino acid, and from the 10th to the 11th. Transition probability, indicating the repeated numbers of state usage, may be given to the self-loop arc as 30/50 = 0.60, and accordingly as 1.0–0.60 = 0.40 to the remaining arc. With these probabilities for the states and arcs, the fitness of the unknown sequence (TKHHH) to the model can be calculated as a likelihood value, which is $1.0 \times 1.0 \times 1.0 \times 1.0 \times 1.0 \times 0.64 \times 0.60 \times 0.64 \times 0.60 \times 0.64 \times 0.40 = 0.038$.

Model training was carried out using the Viterbi algorithm [21], an unsupervised learning algorithm. Baum–Welch algorithm is another learning algorithm commonly used in HMM. Baum–Welch algorithm utilizes more parameters than the Viterbi algorithm for the learning process, which takes more time for its processing, and may involve a local maximum problem depending on the initial estimate of the parameters [22]. Therefore, Baum–Welch algorithm was not used in our model. In the Viterbi algorithm, only the best path with the highest possibility for each sequence is searched and used for the parameter setting during learning. The lone use of the best path provides for a quicker learning process by not using excessive parameters, while maintaining reasonable recognition accuracy.

2.3. Successive state splitting

One of the fundamental problems of HMM is the time-consuming process for the search of an optimal structure for the analysis data. To overcome this problem, we applied the SSS algorithm [23] to estimate an optimal structure of left-to-right HMMs [18]. The basic SSS strategy is to grow the model structure with only one state, from the initial to the final model, with required numbers of states by splitting the state to either parallel or serial direction. The state with maximum Shannon's information entropy [24], indicating the worst characteristic state, which may not represent any feature of the training data, was selected in every splitting step. Furthermore, for each splitting process, every available splitting direction was tried, and the direction that exhibited the highest likelihood values to all the applied sequences was employed.

If every state holds equal distribution of symbol generating probabilities, such model would represent any random sequences with its unrestricted expression ability. The SSS method will save the distinctive states with characteristic symbol generating probability distribution, whereas the non-characteristic states (those with highest entropy) will be targeted for splitting to conserve features underlying in the training data for better prediction model [18].

2.4. Binding prediction scheme

As shown in Fig. 2, the prediction scheme is divided into two steps. First, the prediction model structure is constructed using all data sequences. In the present study, unsupervised learning, which only utilizes high binders sequence data, was used in the construction of the prediction model. In this model construction step, the appropriate prediction model structure was searched by the SSS method.

Secondly, the peptide sequences in the training dataset were applied to the final prediction model for the learning process to rewrite the parameters in the model. After the learning, each peptide in the test dataset was applied to the model to obtain its likelihood value. The likelihood value indicates the fitness of the peptide sequence to the binding tendency in the prediction model conserved through the learning process. Peptides with higher likelihood value indicate higher probabilities of binding to the MHC class II molecule.

2.5. Evaluation of the prediction

As mentioned above, the output of an HMM model is the likelihood value of the peptide tested. Since the absolute value of binding strength cannot be obtained, the peptides tested are compared by means of relative values of likelihood. Peptide that have high likelihood value, relative to that of non-binding peptide, is differentiated from non-binding peptide, and is regarded as a binding peptide. From the feature of HMM, the prediction accuracy of the S-HMM was evaluated using receiver-operating characteristic (ROC) analysis [25].

In ROC analysis, the ROC curve is obtained by plotting true-positive proportion [true-positives/(true-positives + false-negatives)] against false-positive proportion [false-positives/(true-negatives + false-positives)] for various classification thresholds. High-binding and non-binding peptides are designated as positive and negative, respectively. For evaluation of prediction accuracy, the value of area under the ROC curve (AUC) is used as the measure of the predictive performance. An AUC value of 1.0 corresponds to a perfect prediction whereas 0.5 corresponds to prediction by random guessing; empirically, AUC > 0.9 indicates excellent prediction and AUC < 0.7 indicates poor prediction. We considered an AUC value of 0.8 as the threshold for useful predictions. These AUC values provide a universal basis for comparisons among different prediction approaches. To obtain the AUC values, 473 non-binding peptide sequences to HLA-DR1 (Dr. V. Brusic, personal communication) were used as the presumably non-binding data for all the molecules. In the prediction of the discrimination for binding peptides to different allelic molecules, as discussed in Section 3.3, sequences of all peptides that are highly binding to the MHC class II molecule, except the data used for the prediction model construction, were applied in place of the non-binding data.

To evaluate prediction performance using AUC values, threefold (for HLA-DR2, -DR7, -DR17, and -DR51), fourfold (for HLA-DR11, -DR15, and -DQ2), and fivefold (for HLA-DR1 and -DR4) cross validations were performed, depending on the size of the data (Table 2). For example, for

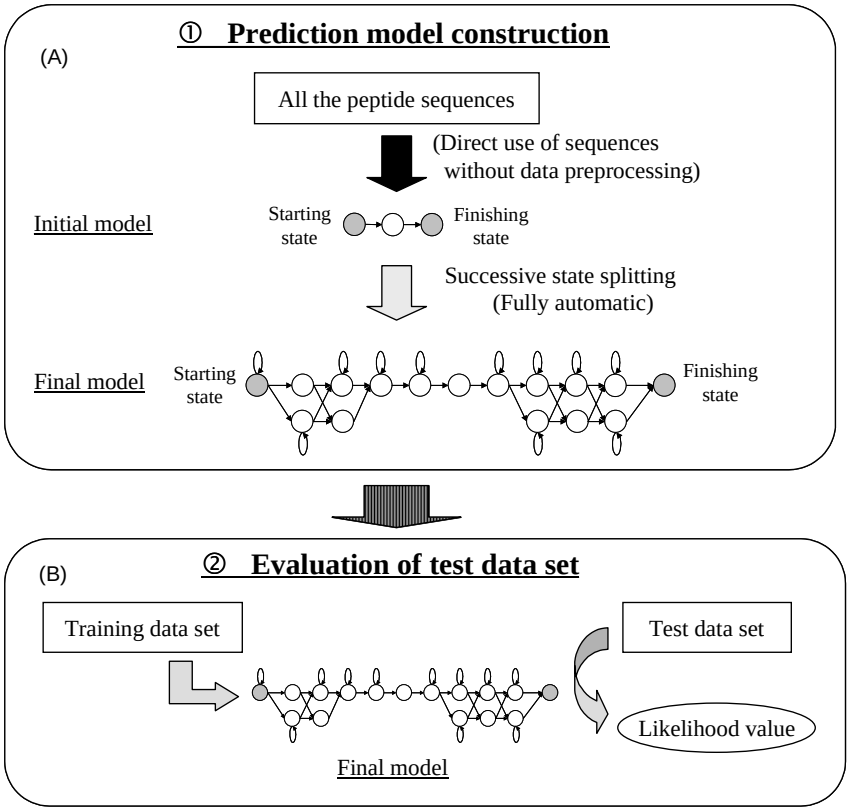

Fig. 2. Schematic diagram of the prediction scheme of S-HMM. In the model construction step, from the initial to the final model using all the peptide data, the optimal prediction structure is automatically searched by SSS (A). In the evaluation step, the likelihood value for each test data is obtained (B). After the model construction, the training dataset is first applied to the final model to rewrite all the parameters reflecting the data. Test dataset is then applied to the model to obtain the likelihood value for every data.

threefold cross validation, the initial peptide data were divided into three subsets such that the distribution of the lengths of binding peptides were uniform in each subsets. Initially, two subsets (nos. 2 and 3) were used as training data, while the remaining subset (no. 1) was applied as test data for the likelihood calculation. Next, subsets nos. 1 and 3 were used as training data, while subset no. 2 was applied as test data. In the same manner, learning and test were carried out using each subset to obtain the AUC values. The average of the AUC values calculated for the cross validation sets was used for the evaluation.

3. Results and discussions

3.1. Features revealing the simplicity of S-HMM prediction

We have shown that the prediction procedure of S-HMM can be completed using two simple steps—prediction model construction and test data evaluation by the likelihood value, as shown in Fig. 2. In the unsupervised learning manner of S-HMM, only the sequences for all binding peptides are required for the prediction, and both quantitative matrixes generated by additional experiments after the binding assays

Table 2
Cross validation results for all the HLA allelic molecules predicted with 15 states model

	DR1 (305)[a]	DR4 (127)[a]	DR11 (67)[a]	DR15 (46)[a]	DQ2 (46)[a]	DR2 (33)[a]	DR7 (34)[a]	DR17 (33)[a]	DR51 (39)[a]
Dataset 1[b]	0.82	0.76	0.90	0.85	1.00	0.89	0.88	0.91	0.84
Dataset 2	0.83	0.83	0.91	0.92	0.91	0.85	0.87	0.85	0.88
Dataset 3	0.82	0.88	0.91	0.79	0.92	0.89	0.82	0.94	0.84
Dataset 4	0.83	0.87	0.93	0.85	100	–	–	–	–
Dataset 5	0.83	0.85	–	–	–	–	–	–	–
Average	0.83	0.84	0.91	0.86	0.96	0.88	0.86	0.90	0.85

[a] The values in parentheses are numbers of peptides selected from the database.
[b] Each data set consists of training data and test data. DR1 and DR4 were divided into five datasets because there were sufficient peptides; DR11, DR15, and DQ2 were divided into four datasets; the remaining data were divided into three datasets.

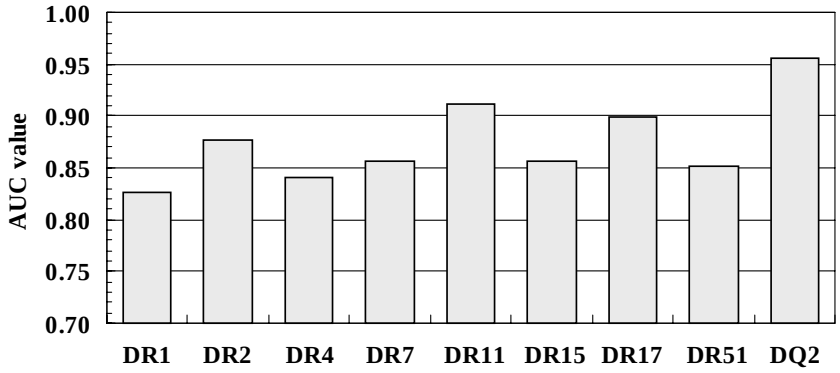

Fig. 3. Comparison of average AUC values from ROC analysis for nine HLA allelic molecules. State number of 15 was used for the prediction.

or the following core-region assumption using the matrix are not required [14,15,17].

The SSS method can construct the prediction model fully automatically, and furthermore, the entire predicting procedure could be calculated within 5 min, even with 1000 peptide sequences, by means of a Pentium III 1 GHz personal computer. Upon the test data evaluation, peptides exhibiting higher likelihood values could be chosen as candidates for further biological assays in the investigations of the actual immune regulatory activity.

3.2. Universal prediction ability of S-HMM

From the ROC analysis, S-HMM showed an average AUC value of 0.87 among the nine different HLA molecules (Fig. 3). Seven prediction models for HLA molecules exhibited high prediction accuracy exceeding AUC value of 0.85. Of note, HLA-DR11 and -DQ2 showed excellent prediction with high average AUC values of 0.91 and 0.96, respectively. To the best of our knowledge, such high AUC values among wide range of MHC class II molecules have not been reported to date. These results imply the universal flexibility and ability of S-HMM toward the screening of peptides interacting with MHC class II molecules.

As shown in Fig. 4, the ROC curves of HLA-DR4, -DR11, and -DQ2 were selected as good examples that show normal, better, and excellent prediction ability, respectively. If the model is perfect, the ROC curve is represented by a left vertical and an upper horizontal axes; in contrast, if the model lacks predictive ability, the ROC curve resembles a diagonal line. As shown in the cross validation results for all MHC class II molecules (Table 2), the AUC values of prediction accuracy were almost identical for each dataset. These results suggest that the database was evenly divided which indicated an equal distribution of peptide lengths. Although our prediction accuracy did not exceed the reported results for HLA-DR1 (AUC = 0.91) [26] and HLA-DR4 (AUC = 0.94) [17], the average AUC value of all nine MHC class II molecules (AUC = 0.87) was considerably high. Such universal high AUC values indicate the high predictive potential of S-HMM for other MHC class II allelic molecules. Furthermore, it is important to note that the HLA-DR11 (AUC = 0.91) and HLA-DQ2 (AUC = 0.96) models showed high AUC values against the other datasets. Considering the simple prediction scheme with features of S-HMM, the resulting above-average AUC values prove that the S-HMM is a useful and universal tool.

In the previous work, the prediction model for HLA-DR1 was constructed using S-HMM fixed with 20 states [18]. To obtain the universal prediction model that can show higher accuracy in all MHC class II molecules, the effects of different state numbers on prediction accuracy were examined (Table 3). State number 15 exhibited the highest average AUC value among the nine MHC class II molecules. The AUC values of HLA-DR7 and -DR15, which lack peptide sequences longer than 20-mer, significantly decreased in prediction models with 20 or 25 states. Such decreases in the prediction ability with larger state number are caused by short peptides, which cannot fit in the long model

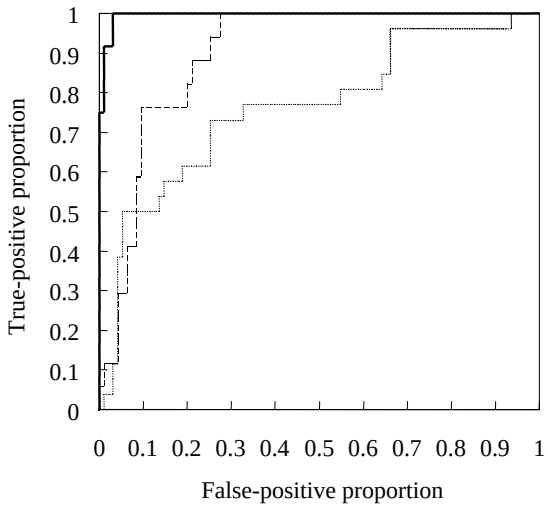

Fig. 4. Some examples of ROC curves obtained from the S-HMM prediction on HLA-DR4, -DR11, and -DQ2. Dotted line, HLA-DR4; broken line, HLA-DR11; solid line, HLA-DQ2.

Table 3
Effect of state number on average AUC value in the prediction of S-HMM

State number	DR1 (305)[a]	DR2 (33)[a]	DR4 (127)[a]	DR7 (34)[a]	DR11 (67)[a]	DR15 (46)[a]	DR17 (33)[a]	DR51 (39)[a]	DQ2 (46)[a]	Average AUC value
5	0.79	0.67	0.85	0.75	0.95	0.86	0.78	0.77	0.91	0.82
10	0.82	0.71	0.84	0.84	0.93	0.93	0.88	0.80	0.96	0.86
15	0.83	0.88	0.84	0.86	0.91	0.86	0.90	0.85	0.96	0.87
20	0.84	0.84	0.86	0.73	0.94	0.57	0.92	0.89	0.97	0.84
25	0.83	0.85	0.85	0.86	0.95	0.50	0.87	0.87	0.93	0.84

[a] The values in parentheses are number of peptides selected from the database.

structure, and result with failure in likelihood calculations. All molecules, except HLA-DR11, showed a decrease in average AUC values in models with state numbers that are less than 10. The decrease of prediction ability with smaller state numbers may be caused by poor specificity, in which the model gives high likelihood values for any sequences, even if non-binding peptide was used. The AUC values of HLA-DR11 and -DQ2 were significantly high, even with different state numbers. In the databases of these two molecules, most of peptides were short peptides (Table 1). A model constructed by such database may possess specific recognition potency. These results have shown that the optimal model state number depends on the distribution of peptide lengths in the objective database, and suggest that the 15-state model functions universally for other MHC class II molecules.

3.3. Discriminative prediction of allelic differences of binding peptides using S-HMM

For the effective screening of peptides that bind to polymorphic molecule, such as MHC class II molecules, the prediction model should discriminate, not only binding from non-binding peptides, but also peptides that bind to the objective allelic molecules from those that bind to other MHC class II molecules. We have carried out studies to examine the latter discriminative recognition ability of S-HMM. To discriminate the S-HMM model for certain allelic molecules, the sequences of high-binding peptides to different HLA molecules were inputted as negative data. For example, after the prediction model is constructed and trained with the binding peptide data of HLA-DR1, the sequences of all binding peptides to HLA-DR2, -DR4, -DR7, -DR11, -DR15, -DR17, -DR51, and -DQ2 were applied individually as negative data. Therefore, the high AUC values can be regarded as "the easiness of discriminating two groups of binding peptides."

As shown in Table 4, five models constructed by different HLA-DR molecules (DR2, DR7, DR15, DR17, and DR51) showed high AUC values in the discriminative prediction between self-binding peptides and DR11-binding peptides. This result indicates that peptides which bind to DR11 were regarded as significantly different from those of five DR molecules. However, a clear tendency to discriminate

Table 4
Average AUC values in the prediction models using high-binding peptides to other HLA molecules that are assumed as non-binding peptides to the model molecule

	Data used as the training data for the construction of the prediction model									Average AUC value among the models
	DR1 (305)[a]	DR2 (33)[a]	DR4 (127)[a]	DR7 (34)[a]	DR11 (67)[a]	DR15 (46)[a]	DR17 (33)[a]	DR51 (39)[a]	DQ2 (46)[a]	
High-binding peptides as test data										
DR1 (305)[a]	–	0.86	0.55	0.85	0.69	0.81	0.87	0.83	0.94	0.80[b]
DR2 (33)[a]	0.62	–	0.61	0.49	0.81	0.77	0.51	0.74	0.85	0.67
DR4 (127)[a]	0.61	0.93	–	0.78	0.86	0.83	0.87	0.74	0.92	0.82
DR7 (34)[a]	0.61	0.48	0.59	–	0.76	0.76	0.51	0.42	0.85	0.62
DR11 (67)[a]	0.44	0.86	0.50	0.87	–	0.86	0.91	0.86	0.92	0.78
DR15 (46)[a]	0.62	0.73	0.63	0.78	0.82	–	0.87	0.67	0.88	0.75
DR17 (33)[a]	0.62	0.53	0.61	0.54	0.81	0.76	–	0.73	0.83	0.68
DR51 (39)[a]	0.61	0.83	0.47	0.64	0.79	0.68	0.75	–	0.77	0.69
DQ2 (46)[a]	0.63	0.84	0.60	0.85	0.82	0.85	0.86	0.84	–	0.79
Average AUC value among the HLA molecules	0.60[c]	0.76	0.57	0.72	0.80	0.79	0.77	0.73	0.87	

[a] The values in parentheses are number of peptides selected from the database.

[b] All average AUC values indicate the average prediction accuracy among the models constructed by different HLA molecules. For the average calculation, the self-AUC value is excluded.

[c] All average AUC values indicate the average prediction accuracy among the HLA molecules. For the average calculation, the self-AUC value is excluded.

differences among other HLA-DR molecules was not observed.

In contrast, the model constructed for HLA-DQ2 binding peptide data showed significantly high AUC values, which exceeded 0.85 against six HLA-DR molecules (HLA-DR1, -DR2, -DR4, -DR7, -DR11, and -DR15). Additionally, AUC values greater than 0.82 were obtained in the predictions discriminating the peptides that specifically bind to HLA-DQ2 molecules using six prediction models constructed by peptides highly binding to different HLA-DR molecules (HLA-DR2, -DR7, -DR11, -DR15, -DR17, and -DR51). These results suggest the clear differences in the binding factors between the peptides that bind to the HLA-DR molecules and those that bind to the HLA-DQ molecules. On the other hand, HLA-DR1 and -DR4 models did not exhibit any discriminative prediction ability for the HLA-DQ2 binding peptides. These observations may indicate the wider acceptance of binding peptides in HLA-DR1 and -DR4 molecules as compared to the other DR molecules. Since the allelic genes belong to a different locus, and the HLA-DR and -DQ molecules show different response on immune diseases, the likelihood differences as revealed by our prediction result is reasonable. Therefore, the possibility of discriminative prediction of binding peptides to different allelic loci by S-HMM, with the use of several models trained by different molecules, was suggested. Such discriminative prediction feature of S-HMM can greatly contribute to the first screening process in selecting appropriate candidates for the objective MHC class II molecule.

3.4. Information from the model structure of S-HMM

Fig. 5 shows the structure of the actual prediction model for HLA-DR4 that was obtained from S-HMM. Arcs that connect every state are designated with the transition probabilities, in which an arc with a higher transition probability is used more frequently, reflecting the training data. The shortest path to the finishing state (state 1 → state 6 → state 11 → state 8 → state 3 → state 4 → state 10 → state 12 → state 9) involving nine states was connected by highly used arcs. This corresponds to the data distribution of HLA-DR4, which contain 47% of 9-mer peptides (Table 1).

The amino acid distribution in each state is also shown as the symbol generating probability in Fig. 5. Tracing the path as mentioned above, the 9-mer binding motif AYAAAATSLA can be assumed by reading the amino acids that are indicated by the highest probability (i.e. symbol generation probability). This motif is frequently extracted from the binding peptides by the model training process. The motif sequence should be regarded as an important binding factor in affecting the binding strength, by partially accommodated in the groove or from outside the groove of the MHC class II molecule. In the S-HMM training which utilizes the entire length of variable sequences for the model training, the overall character of the peptides can be conserved.

In the model structure, the path with significantly high transition probabilities (exceeding 0.84) can be regarded as the shorter motif, commonly existing in the data sequences. The path starting from state 6 to 8 (state 6 → state 11 → state 8) are designated with high transition probabilities on the arcs; furthermore, states 6 and 11 do not possess self-loop transitions. In all three states, the tendency of amino acid usage (i.e. symbol generating probability over 0.20) is clear, and accordingly, for this path, binding motif sequences of YAA and RAA can be assumed. However, such motif is not a solid sequence motif of a 3-mer, since state 8, which represents the third amino acid, has a self-loop transition with a probability of 0.51. With this self-loop state, the 3-mer motif may show different lengths through the addition of another amino acid, which would most likely be an A (alanine) for its higher probability in state 8. However, it is important to restate that such short motif is not defined to be associating with the binding groove, but effects in the binding strength as a result.

3.5. Comparison of prediction of S-HMM with other methods

In previous reports, high prediction accuracies have been reported by using the quantitative matrix itself [12,13], or by using core-region affinity-scores calculated from the quantitative matrix in computational approaches [14,15,17]. However, for prediction methods that utilize quantitative matrixes, time-consuming experimental processes, such as peptide synthesis, purifications of MHC class II molecules or synthesized peptides, mass spectroscopy, Edman degradation, and other processes, were indispensable in obtaining the quantitative matrix [2,27]. In our prediction S-HMM method, only the sequence of peptide data is required for the prediction, and thus quantitative matrixes are unnecessary.

Beside the use of the matrix, both binding and non-binding data for the model construction are also required to attain high accuracy of the prediction methods [14,15,17]. Although binding peptide data exists in several databases that is available on the Internet, peptide sequences that are confirmed as non-binding rarely exist, even in huge and well-assembled databases, such as MHCPEP [19]. Therefore, the unsupervised learning manner of S-HMM, which does not require non-binding peptides, should be an important feature to facilitate effective research work using available data on the Internet. In summary, S-HMM has overcome two major shortcomings of prediction methods reported to date, in allowing high prediction accuracy with its unsupervised learning method.

3.6. Overall potential of the prediction using S-HMM

In the present work herein, binding peptide data from nine MHC class II molecules were applied to construct the prediction model. S-HMM has predicted all the molecules

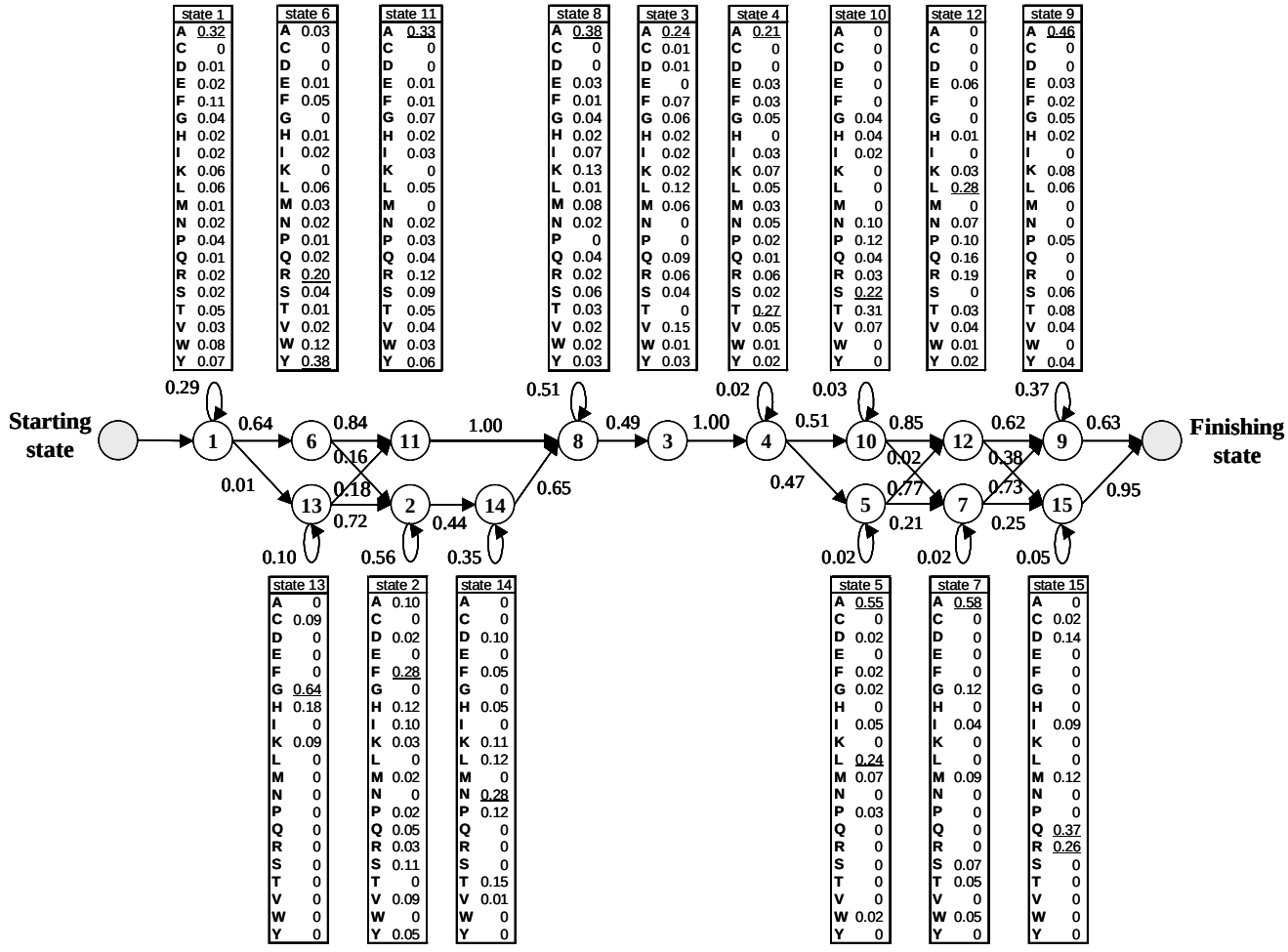

Fig. 5. The actual prediction model structure obtained from S-HMM. In this case, the model is the HLA-DR4 prediction model. Each symbol generating probability distribution is shown as the table above/below the corresponding state. Underlined number in the table highlights the biased generating probability over 0.20. The number associated with each arc is the transition probability.

with high accuracy, with average AUC values exceeding 0.85 when the state number is properly set to 15. S-HMM also showed high accuracy for the discrimination between peptides binding to certain MHC class II molecules from those binding to the other molecules. In the discrimination between HLA-DQ2 and -DR molecule binding peptides, S-HMM showed an average AUC value of 0.87.

We have shown both the simplicity and the potential of our S-HMM as a first screening tool for polymorphic molecules with unclear binding rule, by demonstrating high prediction accuracy using nine different MHC class II molecules, and the discriminative potency of allelic binding difference. In addition to showing universally high prediction accuracy that is comparable to other prediction approaches [17,26], S-HMM has also revealed the great simplicity and application of its prediction scheme. Therefore, it is strongly expected that S-HMM can be utilized for its use in actual research work on other MHC class II molecules.

Acknowledgments

The authors would like to thank Dr. Vladimir Brusic for providing the binding peptide data, and additionally for the personal communications. This study was supported in part by the Grant-in-Aid for Scientific Research (No. 15360439) from the Ministry of Education, Science, Sports, and Culture.

References

[1] Maffei A, Harris PE. Peptide bound to major histocompatibility complex molecules. Peptide 1998;19:179–98.
[2] Ramansee HG, Friede T, Stevanovic S. MHC ligands and peptide motifs: first listing. Immunogenetics 1995;41:178–228.
[3] Mattner F, Fleitmann JK, Lingnau K, Schmidt W, Egyed A, Fritz J, et al. Vaccination with poly-L-arginine as immunostimulant for peptide vaccines: induction of potent and long-lasting T-cell responses against cancer antigens. Cancer Res 2002;62:1477–80.
[4] Jane-wit D, Yu M, Edling AE, Kataoka S, Johnson JM, Stull LB, et al. A novel class II-binding motif selects peptides that mediate

organ-specific autoimmune disease in SWXJ, SJL/J, and SWR/J mice. J Immunol 2002;169:6507–14.
[5] Falk K, Rotschke O. Consensus motifs and peptide ligands of MHC class I molecules. Semin Immunol 1993;5:81–8.
[6] Madden DR. The three dimensional structure of peptide-MHC complexes. Annu Rev Immunol 1995;13:587–622.
[7] Trowsdale J, Lee J, Carvey J, Grosveld F, Bodmer W. Sequences related to HLA-DRa chain on human chromosome 6: restriction enzyme polymorphism detected with DCa chain probes. Proc Natl Acad Sci USA 1983;80:1972–6.
[8] Rundensky AY, Preston-Hurlburt P, Hong SC, Barlow A, Janeway Jr CA. Sequence analysis of peptides bound to MHC class II molecules. Nature 1991;353:622–4.
[9] Brown JH, Jardetzky T, Saper MA, Samraoui B, Bjorkman PJ, Wiley DC. A hypothetical model of the foreign antigen binding site of class II histocompatibility molecules. Nature 1988;332:845–9.
[10] Hammer J, Valsasnini P, Tolba K, Bolin D, Higelin J, Takacs B, et al. Protomiscous and allele-specific anchors in HLA-DR-binding peptides. Cell 1997;74:197–203.
[11] Nelson CA, Petzold SJ, Unanue ER. Peptides determine the lifespan of MHC class II molecules in antigen-presenting cell. Nature 1994;371:250–2.
[12] Marshall KW, Wilson KJ, Lian J, Woods A, Zaller D, Rothbard JB. Prediction of peptide affinity to HLA DRB1*0401. J Immunol 1995;154:5927–33.
[13] Hammer BJ, 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;180:2353–8.
[14] Honeyman MC, Brusic V, Stone NL, Harrison LC. Neural network-based prediction of candidate T-cell epitopes. Nat Biotechnol 1998;16:966–9.
[15] Brusic V, Rudy G, Honeyman M, Hammer J, Harrison L. Prediction of MHC class II-binding peptides using and evolutionary algorithm and artificial neural network. Bioinformatics 1998;14:121–30.
[16] Mallios RR. Class II MHC quantitative binding motifs derived from a large molecular database with a versatile iterative stepwise discriminant analysis meta-algorithm. Bioinformatics 1999;15: 432–9.
[17] Noguchi H, Hanai T, Honda H, Harrison LC, Kobayashi T. Fuzzy neural network-based prediction of motif for MHC class II binding peptides. J Biosci Bioeng 2001;92:227–31.
[18] Noguchi H, Kato R, Matsubara Y, Hanai T, Honda H, Brusic V, et al. Hidden Markov model-based prediction of antigenic peptides that interact with MHC class II molecules. J Biosci Bioeng 2002;94:264–70.
[19] Brusic V, Rudy G, Kyne AP, Harrison L. MHCPEP, a database of MHC-binding peptides: update 1996. Nucleic Acids Res 1997;25:269–71.
[20] Rabiner LR, Juang BH. A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE 1989;77:257–86.
[21] Krogh A, Brown M, Mian IS, Sjoelander K, Haussler D. Hidden Markov models in computational biology. Applications to protein modeling. J Mol Biol 1994;235:1501–31.
[22] Picone JW. Continuous speech recognition using hidden Markov models. IEEE Acoust Speech Signal Process Mag 1990 July;26–41.
[23] Takami J, Sagayama S. Automatic generation of hidden Markov networks by a successive state splitting algorithm. IEICE Trans (in Japanese) 1993;76:2155–64.
[24] Shannon CE. A mathematical theory of communication. Bell Syst Tech J 1948;27:379–423, 623–56.
[25] Swets JA. Measuring the accuracy of diagnostic systems. Science 1998;24:1285–93.
[26] Brusic V, Zeleznikow J, Sturniolo T, Bono E, Hammer J. Data cleansing for computer models: a case study from immunology. In: Proceedings of ICONIP99, Sixth International Conference on Neural Information Processing. IEEE 1999;603–9.
[27] Sinigaglia F, Hammer J. Defining rules for the peptide-MHC class II interaction. Curr Opin Immunol 1994;6:52–6.