BIOINFORMATICS

Vol. 20 no. 9 2004, pages 1388–1397
DOI: 10.1093/bioinformatics/bth100

Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach

Morten Nielsen[1,*], Claus Lundegaard[1], Peder Worning[1], Christina Sylvester Hvid[2], Kasper Lamberth[2], Søren Buus[2], Søren Brunak[1] and Ole Lund[1]

[1]Center for Biological Sequence Analysis, BioCentrum-DTU, Building 208, Technical University of Denmark, DK-2800 Lyngby, Denmark and [2]Department of Experimental Immunology, Institute of Medical Microbiology and Immunology, Panum Building 18.3.22, University of Copenhagen, Blegdamsvej 3B, DK-2200 Copenhagen N, Denmark

Received on August 29, 2003; revised on November 21, 2003; accepted on December 9, 2003
Advance Access publication February 12, 2004

ABSTRACT

Motivation: Prediction of which peptides will bind a specific major histocompatibility complex (MHC) constitutes an important step in identifying potential T-cell epitopes suitable as vaccine candidates. MHC class II binding peptides have a broad length distribution complicating such predictions. Thus, identifying the correct alignment is a crucial part of identifying the core of an MHC class II binding motif. In this context, we wish to describe a novel Gibbs motif sampler method ideally suited for recognizing such weak sequence motifs. The method is based on the Gibbs sampling method, and it incorporates novel features optimized for the task of recognizing the binding motif of MHC classes I and II. The method locates the binding motif in a set of sequences and characterizes the motif in terms of a weight-matrix. Subsequently, the weight-matrix can be applied to identifying effectively potential MHC binding peptides and to guiding the process of rational vaccine design.

Results: We apply the motif sampler method to the complex problem of MHC class II binding. The input to the method is amino acid peptide sequences extracted from the public databases of SYFPEITHI and MHCPEP and known to bind to the MHC class II complex HLA-DR4(B1*0401). Prior identification of information-rich (anchor) positions in the binding motif is shown to improve the predictive performance of the Gibbs sampler. Similarly, a consensus solution obtained from an ensemble average over suboptimal solutions is shown to outperform the use of a single optimal solution. In a large-scale benchmark calculation, the performance is quantified using relative operating characteristics curve (ROC) plots and we make a detailed comparison of the performance with that of both the TEPITOPE method and a weight-matrix derived using the conventional alignment algorithm of ClustalW. The calculation demonstrates that the predictive performance of the Gibbs sampler is higher than that of ClustalW and in most cases also higher than that of the TEPITOPE method.

Contact: mniel@cbs.dtu.dk

INTRODUCTION

Only a small fraction of the possible peptides that can be generated from proteins of pathogenic organisms actually generate an immune response. In order to be presented to CD8+ cytotoxic T lymphocytes (CTLs), a precursor peptide must be generated by the proteasome and transported to the endoplasmic reticulum (ER) by the transporter associated with antigen processing (TAP) before it can bind to a major histocompatibility complex class I (MHC I) molecule (Serwold *et al*., 2002). Hereafter, it is transported to the cell surface, where it may induce a CTL response that can kill infected cells. Whereas MHC I molecules mainly sample peptides from the cytosol, MHC II samples peptides derived from endocytosed proteins. Unfolded polypeptides bind MHC II in the endocytic organelles [reviewed by Castellino *et al*. (1997)]. Peptides presented by MHC II in turn activate CD4+ helper T lymphocytes (HTLs) to stimulate cellular and humoral immunity against the appropriate microorganism.

The most selective step in antigen presentation is the binding to the MHC molecule (Yewdell *et al*., 1999). The specificity of this binding and that of some of the other processes involved in antigen presentation can be predicted from the amino acid sequence. Such predictions can be used to select epitopes for use in rational vaccine design and to increase the understanding of the role of the immune system in infectious diseases, autoimmune diseases and cancers.

[*] To whom correspondence should be addressed.

Both MHC I and MHC II are highly polymorphic (variable at each gene locus), and the specificity of the alleles (variants) is often very different. Thus different individuals will typically react to a different set of peptides from a given pathogen. This high polymorphism has large implications for vaccine design. Each of the MHC molecules has a potential different specificity, and if a vaccine needs to contain a unique peptide for each of these molecules, it will need to comprise thousands of peptides. Moreover, the task of deriving MHC prediction algorithms would be immense. Nevertheless, many MHC alleles have very similar binding specificities, and it is therefore often possible to find promiscuous peptides that bind to a series of MHC variants. This has two important implications. First, it limits the number of epitopes needed to be included in a vaccine design. Second, it allows high accuracy predictions for MHC alleles also in situations where the binding motif is poorly characterized (Brusic *et al*., 2002)

The core binding motif of both MHC I and MHC II is approximately nine amino acids long (Rammensee *et al*., 1999, http://www.uni-tuebingen.de/uni/kxi/). Whereas the peptide binding groove in the MHC I molecule tends to be closed at either end and MHC I rarely binds peptides much longer than nine amino acids, the ends of the MHC II binding groove are open. Consequently, MHC class II can accommodate much longer peptides—possibly even whole proteins (Sette *et al*., 1989a; Castellino *et al*., 1997). This difference has important implications for the development of algorithms predicting binding. The specificity of an MHC I molecule can be derived by extracting the motif from a set of 9mer peptides known to bind to a given allele. In contrast, a set of peptides binding MHC II will typically be of different length, and therefore they need to be aligned correctly before the nine amino acid long core-binding motif can be identified.

Many different methods have been applied to predict peptide–MHC binding, including simple binding motifs, quantitative matrices, hidden Markov models and artificial neural networks. For class I, these gap- and alignment-free methods can readily be applied since the binding motif is well characterized and most natural peptides that bind MHC class I are of close to equal length (Parker *et al*., 1994; Brusic *et al*., 1994; Rammensee *et al*., 1999; Buus *et al*., 2003; Nielsen *et al*., 2003). However, the situation for MHC class II binding is quite different due to the great variability in the length of natural MHC binding peptides. This length variability makes alignment a crucial and integrated part of estimating the MHC binding motif and predicting peptide binding. Quantitative matrices estimated from experimentally derived position specific binding profiles have given reasonable performance in prediction of MHC class II binding (Sette *et al*., 1989b; Hammer *et al*., 1994; Marshall *et al*., 1995; Sturniolo *et al*., 1999). However, such matrices are very costly to derive, and more importantly they lack the flexibility of data driven machine-learning methods to be refined in an iterative manner when more data become available. Brusic *et al*. (1998a) have described a hybrid method for predicting peptide–MHC class II binding. They handle the alignment problem using an evolutionary algorithm and subsequently apply artificial neural networks to classify peptides as binding/non-binding.

The advanced motif sampler method described here is based on the Gibbs sampling method described by Lawrence *et al*. (1993). The Gibbs sampler method has earlier been used extensively for location of transcription factor binding sites (Thompson *et al*., 2003) and in the analysis of protein sequences (Lawrence *et al*., 1993; Neuwald *et al*., 1995). The method attempts to find an optimal local alignment of a set of N sequences by means of Monte Carlo Metropolis (Metropolis *et al*., 1953) sampling of the alignment space. The scoring-function guiding the Monte Carlo search is defined in terms of fitness (information content) of a log-odds matrix calculated from the alignment. The general problem to be solved by the motif sampler is to locate and characterize a pattern contained within a set of N amino acids (or DNA) sequences. In situations where the sequence pattern is very subtle and the motif weak, this is a highly complex task, and conventional multiple sequence alignment programs will typically fail. In the following, we describe a novel implementation of the Gibbs sampler method specialized and optimized to locate and characterize the motif of MHC class I and class II binding. The method applies the advanced techniques of sequence weighting and pseudo-count correction for low counts as well as differential position specific weighting (C. Lundegaard *et al*., unpublished data) and generation of consensus weight-matrices to estimate the binding motifs.

MATERIALS AND METHODS

Weight-matrix calculation

A central part of the motif sampler algorithm is the weight-matrix calculation, including pseudo-count correction for low counts and sequence weighting, from a given sequence alignment. In order to optimize pseudo-count correction and sequence weighting parameters of the prediction method, we first use the Gibbs sampler to calculate weight-matrices from pre-aligned sequences restricted to a series of MHC class I molecules.

MHC class I binding data

Peptides known to bind MHC class I molecules were extracted from the databases of SYFPEITHI (Rammensee *et al*., 1999, http://syfpeithi.bmi-heidelberg.com/) and MHCpep (Brusic *et al*., 1998b, http://wehih.wehi.edu.au/mhcpep/). Only peptides of length 9 were included. The peptides were clustered into the nine supertypes (A1, A2, A3, A24, B7, B27, B44, B58 and B62) as described by Sette and Sidney (1999). These peptides constitute the training set for the MHC class I binding weight-matrices. Datasets of peptides for which the binding affinity to the MHC had been measured as described by Sylvester-Hvid *et al*. (2002) were available to us for four of

Table 1. Data for the training and evaluation of the HLA class I binding predictions

Supertype	N	Allele	N	N_{bind}
A1	92	HLA-A*0101	283	27
A2	626	HLA-A*0204	528	144
A3	228	HLA-A*0301	212	5
B7	201	HLA-B*0702	154	24

The first column gives the supertype names included in the calculation, the second column the number of unique 9mer peptides in the training set, the third column the HLA allele name for the evaluation set data, and the fourth and fifth columns the total number of peptides and the number of binders in the evaluation set, respectively. Binders were determined using a threshold of 500 nM.

the nine supertypes (A1, A2, A3 and B7). These datasets were used to evaluate the prediction accuracy of the corresponding weight-matrix. To avoid over-training, any peptide found in the training set was removed from the evaluation set. In Table 1 we provide the number of unique peptides in the training set, the number of peptides in the evaluation set, the corresponding allele names and the number of binding peptides (affinity stronger than 500 nM) for each of the four supertypes, respectively.

Optimization of parameters for deriving amino acid frequencies

The Gibbs sampler has a series of free parameters defining how a weight-matrix is calculated from a multiple alignment. The optimal parameter setting for the parameters is determined in a large-scale benchmark calculation using the above datasets. The most important parameters are

(i) Sequence weighting method.
(ii) Null model.
(iii) Pseudo-count correction method.
(iv) Weight on pseudo-count correction.
(v) Position specific weighting.

(i) Two different strategies for sequence weighting were tested: sequence clustering and sequence weighting as described by Henikoff and Henikoff (1994). For sequence clustering, we use a Hobohm 1-like (Hobohm *et al*., 1992) algorithm with ungapped alignment and sequence identity of 62% as cluster threshold. After the clustering, each peptide in a cluster is assigned a weight equal to $1/N_c$, where N_c is the cluster size. In the Henikoff and Henikoff sequence weighting scheme an amino acid is assigned a weight $w = 1/rs$, where r is the number of different amino acids at a given position in the alignment and s the number of occurrences of the amino acid. The weight of a sequence is then assigned as the sum of the amino acid weights. The method of Henikoff and Henikoff is fast as the computation time only increases linearly with the number of sequences. For the clustering algorithm on the other hand, the computation time increases as the square of the number of sequences.

(ii) To estimate the significance of a given alignment, the Gibbs sampler compares the information content with a null model. The null model is defined in terms of background amino acid frequencies. Three distinct null models were tested: the amino acid distribution in the SWISS-PROT database (Bairoch and Apweiler, 2000), a flat distribution and an amino acid distribution estimated from the raw counts of the peptides in the input to the Gibbs sample.

(iii) Two strategies for pseudo-count correction were tested: Equal and Blosum correction. In both cases the pseudo-count frequency is estimated as described by Altschul *et al*. (1997). For the Equal correction, a substitution matrix with identical frequencies for all amino acid substitutions is applied. For Blosum correction, a Blosum62 (Henikoff and Henikoff, 1992) substitution matrix is applied.

(iv) The effective amino acid frequency is calculated as (Altschul *et al*., 1997)

$$f = \frac{\alpha \cdot f' + \beta \cdot g}{\alpha + \beta}.$$

Here f' is the observed frequency, g the pseudo-count frequency, α the effective sequence number and β the weight on the pseudo-count correction. When the sequence weighting is performed using clustering, the effective sequence number is equal to the number of clusters. When sequence weighting as described by Henikoff and Henikoff (1992) is applied, the mean number of different amino acids in the alignment gives the effective sequence number.

(v) In many situations, prior knowledge about the importance of the different positions in the binding motif exists. Such prior knowledge can be included with success in the search for binding motifs (Lundegaard *et al*., unpublished data).

Gibbs sampling

The algorithm samples possible alignments of the N sequences. For each alignment a log-odds weight matrix is calculated as $\log(p_{i,j}/q_i)$, where $p_{i,j}$ is the frequency of amino acid i at position j in the alignment and q_i the background frequency of that amino acid. The values of $p_{i,j}$ are estimated using sequence weighting and pseudo-count correction for low counts. Sequence weighting is performed using either the method described by Henikoff and Henikoff (1994) or a clustering algorithm, and the pseudo-count correction for low counts is performed using a Blosum weighting scheme (Altschul *et al*., 1997).

The fitness (energy) of an alignment is calculated as

$$E = \sum_{i,j} p_{ij} \cdot \log \frac{p'_{ij}}{q_i}, \qquad (1)$$

where p_{ij} is the occupancy number of amino acid i at position j in the alignment, p'_{ij} is the pseudo-count and sequence weight corrected amino acid frequency of amino acid i at position j in the alignment. Finally, q_i is the background frequency of amino acid i. E is equal to the sum of the relative entropy or the Kullback–Leibler distance (Kullback and Leibler, 1951) in the window.

The set of possible alignments is even for a small dataset very large. For a set of 50 peptides of length 10, the number of different alignments with a core of nine amino acids is $20^{50} \sim 10^{15}$. This number is clearly too large to allow a sampling of the complete alignment space. Instead, we apply the Metropolis Monte Carlo algorithm (Metropolis *et al.*, 1953) to perform an effective sampling of the alignment space.

Two distinct Monte Carlo moves are implemented in the algorithm: (1) single sequence move and (2) phase shift move. In the single sequence move, a new starting point for the alignment of a sequence is selected at random. In the phase shift move, the entire alignment is shifted a random number of residues to the left or right. This last move allows the program to escape efficiently local minima. This may for example occur if the window overlaps the most informative motif but is not centred on the most informative pattern.

The probability of accepting a move in the Monte Carlo sampling is defined as

$$P = \min\left[1, \exp\left(\frac{dE}{T}\right)\right], \tag{2}$$

where dE is the difference in energy between the end and start configurations and T a scalar. Note that we seek to optimize the energy function; hence the positive sign for dE in the equation. T is a scalar that is lowered during the calculation. Equation (2) gives us the result that moves that increase E will always be accepted ($dE > 0$). On the other hand, only a fraction, given by $e^{dE/T}$, of the moves that decrease E will be accepted. For high values of the scalar $T (T \gg dE)$, this probability is close to 1; however, as T is lowered during the calculation, the probability of accepting unfavorable moves will be reduced, forcing the system into a state of high fitness (energy).

MHC class II binding data

We extracted peptides binding to the MHC class II molecule HLA-DR4(B1*0401) from the SYFPEITHI (Rammensee *et al.*, 1999) and MHCPEP (Brusic *et al.*, 1998a) databases. The dataset consists of 532 unique peptides sequences. Peptides that did not allow a hydrophobic residue at the P1 position in the binding motif were removed (Brusic *et al.*, 1998a). That is, a peptide was removed if no hydrophobic residues were present at the first $N - L + 1$ positions, where N is the peptide length and L is the motif length. The hydrophobic filter removed 28 peptides. Furthermore, the dataset was reduced to remove unnatural peptide sequences with an extreme amino acid content by removing peptides with more than 75% alanine. The final training set had 456 unique peptides. The length distribution in the training set ranges from 9 to 30 residues, with the majority of peptides having a length of 13 amino acids.

HMMER weight-matrices

To evaluate the performance of the Gibbs sampler method, we estimate the amino acid frequencies and corresponding weight-matrices using the HMMER (Eddy, 1998) package program `hmmbuild` with the following command line options: `--fast --pam BLOSUM62`; here Blosum62 is the Blosum62 score matrix (Henikoff and Henikoff, 1992).

RESULTS

Weight-matrix parameter settings from aligned class I binding peptides

We applied the Gibbs sampler to the MHC class I binding motif problem in order to estimate the optimal setting for the parameters that determine the generation of weight-matrices from fixed alignments. For each parameter setting, we estimate weight-matrices for the four supertypes A1, A2, A3 and B7 using the peptides in the training sets and subsequently evaluate the predictive performance on the corresponding evaluation set. The predictive performance is calculated using both the Pearson correlation coefficient between the log-transformed affinities and the weight-matrix predictions (Nielsen *et al.*, 2003), and the non-parametric A_{roc} measures [the area under the relative operating characteristics (ROC) curve (Swets, 1988)]. By applying the same parameter setting to all four supertypes, we minimize the risk of over-fitting. As a comparison, we evaluate the predictive performance of weight-matrices derived using the HMMER package (see Materials and methods section) on the four evaluation sets.

In all situations, the use of a Blosum62 matrix (Blosum correction) for estimating the pseudo-counts gave better predictive performance than using an equal matrix (Equal correction). The background distribution of amino acids estimated from SWISS-PROT also gave a consistently higher predictive performance than those of both a flat distribution and a distribution estimated from the sequence input to the algorithm. In the rest of the analysis, we hence use the Gibbs sampler with the pseudo-count estimated as Blosum correction and the background amino acid distribution estimated from the SWISS-PROT database.

Figure 1 shows the prediction accuracy estimated in terms of the Pearson correlation coefficient and the A_{roc} value, respectively, for the two different sequence-weighting schemes for a series of pseudo-count weights (β) for four supertypes. As a comparison, the prediction accuracy of the weight-matrices estimated using HMMER as well as the prediction accuracy using the SYFPEITHI prediction method is shown. It is clear that the two sequence-weighting schemes have similar predictive performance and that the optimal performance is found for a value of the pseudo-count weight β close to 50 for the Henikoff and Henikoff (1994) sequence weighting and for

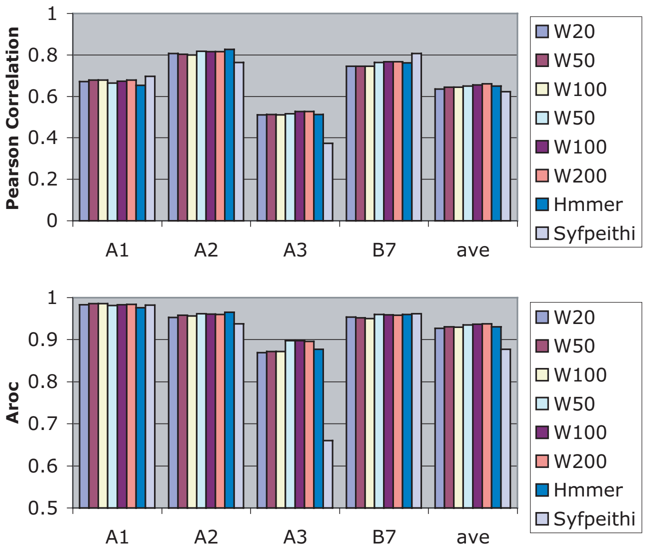


Fig. 1. Predictive performance of the Gibbs sampler for the two schemes of sequence weighting of Henikoff and Henikoff and sequence clustering, respectively. The figure compares the predictive performance in terms of the Pearson correlation coefficient (upper plot) and A_{roc} (lower plot) for the four supertypes A1, A2, A3 and B7 as well as the average of the four. The ROC curves were calculated using a threshold of 500 nM to define binders/non-binders. For the Henikoff and Henikoff sequence-weighting scheme, the performance is given for pseudo-count weights of 20, 50 and 100. For clustering, performance is shown for pseudo-count weights of 50, 100 and 200, respectively. For each supertype, the last two columns give the performance of the HMMER package and the SYFPEITHI Web site predictor, respectively.

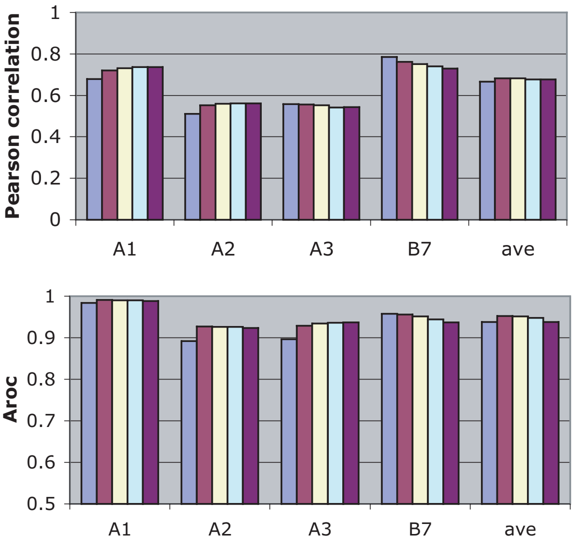


Fig. 2. Prediction performance of the Gibbs sampler for different position specific weight values. The upper figure gives the performance in terms of the Pearson correlation and the lower figure the A_{roc} values for a relative weight of 1, 2, 3, 5 and 9, respectively, on the selected positions. The ROC curves were calculated as described in Figure 1. The last set of bars in each figure give the average performance over the four supertypes.

a value close to 200 for the clustering sequence weighting, respectively. Since the sequence-weighting scheme, based on sequence clustering, has slightly better performance, we will in the following use this sequence-weighting scheme, and consequently we set the pseudo-count weight to 200. Moreover, from the table it is clear that the predictive performance of the Gibbs sampler is comparable with those of both HMMER and the SYFPEITHI prediction method.

As stated previously, prior knowledge regarding the importance of the different positions in the binding motif exists. This is for example the case for the MHC class I binding motif, where the binding for most alleles is largely determined by the fitness of the peptide to the binding pockets at positions 2 and 9 in the motif. Such prior knowledge can be included in the search for binding motifs (Lundegaard *et al.*, unpublished data). Figure 2 shows the predictive performance of the weight-matrix for class I binding when such position specific weighting is included in the motif search. The position specific weighting scheme is determined as the set of anchor residues defined in the SYFPEITHI database, extended with auxiliary anchors if they occur at position 2 or 9. For the A1 supertype, positions 3 and 9 are specified as anchor positions, whereas positions 2 and 7 are auxiliary anchor positions. This means that positions 2, 3 and 9 are included as positions with high weight in the motif search for this supertype. For the other supertypes of A2, A3 and B7, the motif positions with higher weight are positions 2 and 9.

From the results shown in the table, it is clear that a position specific weighting of 2–3 gives the highest predictive performance.

Alignment of MHC class II binding peptides and weight-matrix extraction

We applied the Gibbs sampler to estimate the binding motif and corresponding weight-matrix for the HLA-DR4(B1*0401) molecule. We applied the Gibbs sampler with the parameter settings described earlier. In order to ensure that only hydrophobic residues were present at the P1 position in the motif, we restricted the single sequence move in the Monte Carlo procedure to only select from the set of hydrophobic amino acids. The scalar T was initialized to 0.15 and lowered to 0.001 in 10 uniform steps. At each value of T, 5000 Monte Carlo moves were performed. The acceptance of a move was determined using Equations (1) and (2). The motif length was fixed at nine amino acids. The alignment space has a very large number of local maxima with close to identical energy. In order to achieve an effective sampling of these local maxima, we repeated 100 MC calculations with different initial configurations. In Figure 3, we show the

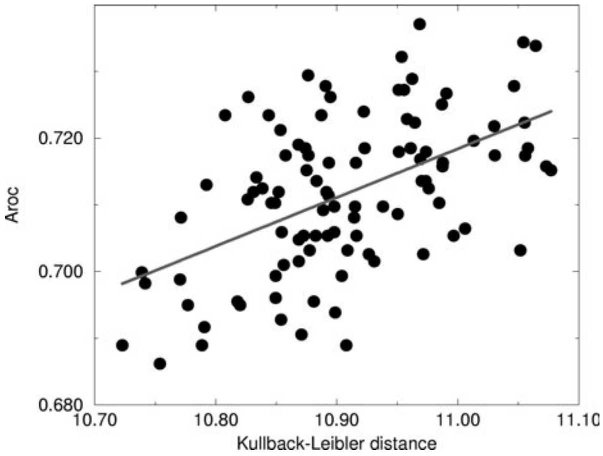


Fig. 3. Predictive performance as a function of the Kullback–Leibler distance. A total of 100 weight-matrices were estimated from distinct Monte Carlo calculations. The different weight-matrices were evaluated on the set of 105 peptides described in the text, and the predictive performance in terms of an A_{roc} value is plotted as a function of the Kullback–Leibler distance. A least-square straight-line fit is shown in red. The correlation coefficient is 0.53.

predictive performance for the 100 weight-matrix solutions as a function of the Kullback–Leibler distance [Equation (1)] estimated from the final sequence alignment. The predictive performance is evaluated on a set of 105 peptides described by Geluk *et al*. (1998).

Figure 3 demonstrates clearly that the Kullback–Leibler distance correlates to some extent with the predictive performance. However, the correlation is not perfect, and the optimal solution is not the one with the optimal predictive performance. Wanting to obtain an effective sampling of the suboptimal solutions, we calculated a consensus weight-matrix as the average over the top five highest scoring weight-matrices (averages over five, 10 and 20 top scoring matrices, respectively, gave similar results).

From the SYFPEITHI database, the anchor positions in the binding motif were estimated to be located at positions 1, 4, 6, 7 and 9, respectively. Anchor positions estimated from a logo-plot of a weight-matrix calculated using the Gibbs sampler with equal weights on all positions confirmed this weighting scheme at all positions except position 7 (Fig. 4, right panel). Hence, we used positions 1, 4, 6 and 9 with an increased weight to guide the Gibbs sampling.

As an estimate of how other conventional alignment methods perform on the motif detection problem, we aligned the sequences in the training set using the ClustalW package (Chenna *et al*., 2003) with a high gap opening penalty to ensure ungapped alignment since initial experiments showed that this resulted in the best performance. Furthermore, we generated a control by placing the sequences in a random alignment with hydrophobic amino acids at the P1 position. From the alignments, we estimated the amino acids frequencies in the nine amino acid long core-region and made logo-plots from these frequency estimates (Fig. 4).

Figure 4 demonstrates that the identification of the binding motif from the training data is indeed a complex and difficult task. The ability of the Gibbs sampler method to detect the subtle sequence motif in a set of peptide sequences is apparent from the figure. The algorithm of ClustalW is on the other hand unable to detect any motif signal except from the strong hydrophobic amino acid preference at position P1. In Figure 5, we show a part of the alignment obtained by the Gibbs sampler for the HLA-DR4(B1*0401) binding motif recognition. Figures 4 and 5 demonstrate how the Gibbs sampler, through the Monte Carlo moves, is able to place the sequences in a register and move from an initial random configuration with close to zero information content to a final alignment configuration with high information content describing the peptide binding motif in detail.

Benchmark calculations

The predictive performance of the Gibbs sampled weight-matrix was benchmarked on 10 datasets and compared with that of the TEPITOPE method (Sturniolo *et al*., 1999) as well with the weight-matrix derived from the ClustalW alignment. The 10 datasets are the eight datasets described by Raghava (MHC-Bench, http://www.imtech.res.in/raghava/mhcbench) and two experimental datasets described by Southwood *et al*. (1998) and Geluk *et al*. (1998). The binding of a peptide was calculated as the score of the highest scoring 9mer subpeptide. We used the non-parametric A_{roc} measure (Swets, 1988) to compare the accuracy of the different prediction methods. In order to calculate a ROC curve, one must classify the dataset into binders and non-binders. For the eight MHCbench datasets, peptides with an associated binding value of zero were assigned to be non-binding, and all other peptides were binders. For the datasets of Southwood and Geluk, datasets, an affinity of 1000 nM was taken as the threshold for peptide binding (Southwood *et al*., 1998) (similar results were obtained for threshold values in the range 500–10 000 nM). Wanting to reduce the chance of over-fitting by evaluating the prediction performance on data points included in the training, we repeated the benchmark calculation on homology-reduced datasets. The homology reduction was performed so that no data point in the evaluation sets had a match in the training set with sequence identity >90% over an alignment length of at least nine amino acids. Table 2 gives a brief description of both the original and the homology-reduced benchmark datasets in terms of the number of peptides and the number of binders, respectively.

In Figure 6A we show the results of the benchmark calculation. From the figure it is clear that the Gibbs sampled weight-matrix has comparable or better predictive performance than those of both TEPITOPE and the ClustalW weight-matrix. In all cases, the ClustalW weight-matrix has a performance that is lower than that of the Gibbs sampled matrix. In order

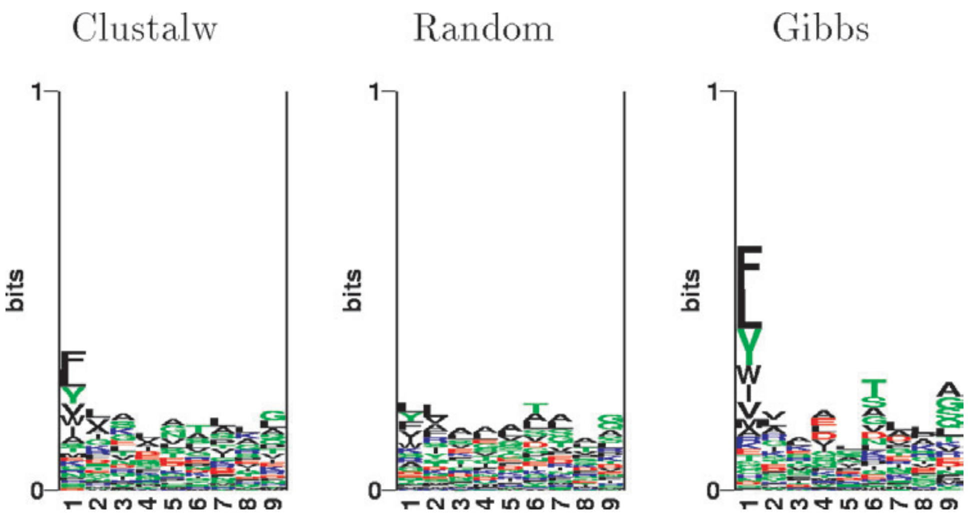
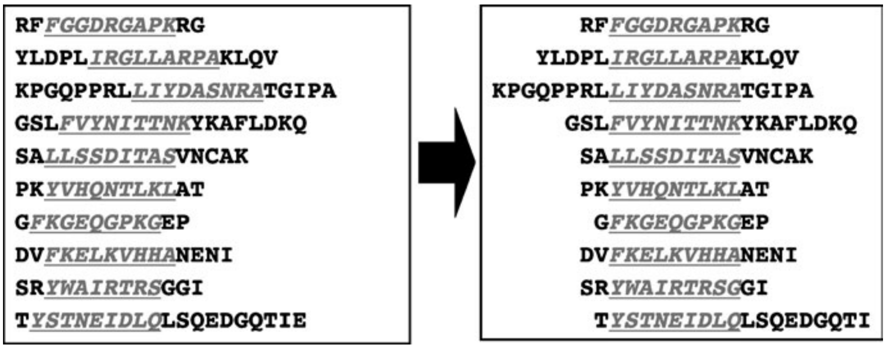


Fig. 4. Logo-plots of amino acid frequencies in three distinct alignments of the peptides in the training set. The alignments were performed using the methods of ClustalW, a random placement and the Gibbs motif sampler, respectively. The height of a column in the logo is proportional to the information content in the sequence motif, and the letter height is proportional to the amino acid frequency (Schneider and Stephens, 1990).


RF*FGGDRGAPK*RG
YLDPL*IRGLLARPA*KLQV
KPGQPPRL*LIYDASNRA*TGIPA
GSL*FVYNITTNK*YKAFLDKQ
SA*LLSSDITAS*VNCAK
PK*YVHQNTLKL*AT
G*FKGEQGPKG*EP
DV*FKELKVHHA*NENI
SR*YWAIRTRS*GGI
T*YSTNEIDLQ*LSQEDGQTIE

RF*FGGDRGAPK*RG
YLDPL*IRGLLARPA*KLQV
KPGQPPRL*LIYDASNRA*TGIPA
GSL*FVYNITTNK*YKAFLDKQ
SA*LLSSDITAS*VNCAK
PK*YVHQNTLKL*AT
G*FKGEQGPKG*EP
DV*FKELKVHHA*NENI
SR*YWAIRTRSG*GI
T*YSTNEIDLQ*LSQEDGQTI

Fig. 5. An alignment generated by the Gibbs sampler for the DR4(B1*0401) binding motif. In the left panel are shown the unaligned sequences, and in the right panel the aligned sequences. The core motif is shown underlined and in italic.

to estimate the significance of the difference in performance between the Gibbs sampler and the TEPITOPE methods, we performed a bootstrap experiment (Press *et al.*, 1993). For each of the datasets, we generate $M = 1000$ datasets by extracting N data points with replacement. Here, N is the number of data points in the original dataset. The performance of both the Gibbs and the TEPITOPE methods was evaluated on each of the datasets, and the p-value for the hypothesis that the TEPITOPE method performs better than the Gibbs sampler is estimated as the fraction of experiments where TEPITOPE has the better performance of the two. The results of this calculation demonstrated that for 5 of the 10 datasets (the Southwood set, set 1, set 2, set 4A and set 4B) the Gibbs sampler method had a performance that is significantly higher than that of TEPITOPE ($p < 0.05$). Only for one dataset (set 5B), did the TEPITOPE method perform better than the Gibbs sampler ($p = 0.96$). For the remaining four datasets, the difference in predictive performance was found to be insignificant ($0.05 < p < 0.95$).

The average A_{roc} values for the Gibbs sampled matrix, the TEPITOPE matrix and the ClustalW matrix methods are 0.744, 0.702 and 0.667 for the complete dataset and 0.673, 0.630 and 0.599 for the reduced datasets, respectively.

For 2 of the 10 datasets (set 5A and set 5B), the TEPITOPE weight-matrix had a higher A_{roc} value than the Gibbs matrix. For the set 5B, this difference is statistically significant ($p = 0.96$). In order to analyze why the Gibbs sampled weight-matrix has poor performance on the two datasets, we estimated the amino acid composition in the two sets as compared with that of the other benchmark sets and the training set. In this analysis we found that both sets had an extremely high content of cysteines in the subset of peptides that bind MHC. In set 5B, for instance, 45 of the 85 peptides contain at least one cysteine, and 37 of these 45 bind MHC. These numbers stand in contrast

Table 2. Description of the MHC class II benchmark datasets

Set	Original		Homology-reduced	
	N	N_b	N	N_b
MHCbench 1	1017	694	496	226
MHCbench 2	673	381	416	161
MHCbench 3A	590	373	334	130
MHCbench 3B	495	279	325	128
MHCbench 4A	646	323	381	111
MHCbench 4B	584	292	375	120
MHCbench 5A	117	70	110	65
MHCbench 5B	85	48	84	47
Southwood	22	16	21	15
Geluk	105	22	99	19

The first column gives the name of the dataset, the second and third columns the number of peptides and the number of peptides classified as binders for the complete sets, the fourth and fifth columns the same numbers for the reduced datasets, respectively. For the Southwood and Geluk datasets a threshold of 1000 nM and for the MHC-bench datasets a threshold value of 0.5, was used to determine binders.

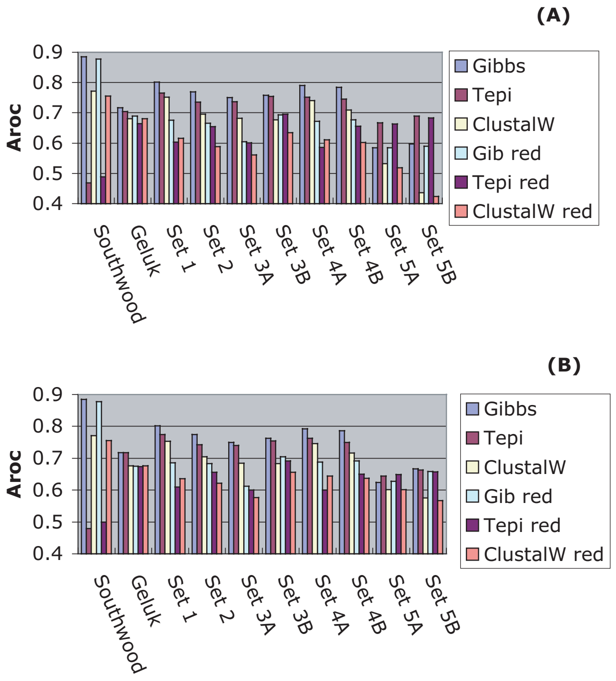


Fig. 6. Prediction accuracy of the Gibbs sampled, the TEPITOPE and the ClustalW weight-matrix methods, respectively, for the datasets described in the text. (**A**) The result for the original data. (**B**) Cysteine substituted benchmark where all occurrences of cysteines are replaced with alanine. For each dataset, the first three bars give the performance on the complete datasets, and the last three bars the performance on the reduced datasets, respectively.

to the low cysteine content in the training set. Here, only 47 of the 456 peptide sequences contain cysteine. The TEPITOPE weight-matrix has a particular behavior for cysteines in that the score for this amino acid at all positions is zero. To verify whether the cysteine content could explain the poor behavior of the Gibbs sampled matrix method as compared with the TEPITOPE matrix method, we repeated the above benchmark calculation substituting all occurrences of cysteine to alanine in the benchmark datasets. The result of the calculation is shown in Figure 6B.

From Figure 6B, it is clear that the Gibbs sampled weight-matrix in the cysteine substituted benchmark calculation for the reduced datasets also has better or comparable predictive performance compared with that of the TEPITOPE matrix method. In particular, one should note that the performance on the two sets 5A and 5B is comparable for the two methods. Repeating the bootstrap experiment for set 5B applying cysteine substitution gave a p-value of 0.5. This demonstrates that it indeed was the unusual cysteine content that led to the poor performance of the Gibbs sampler method for the two datasets. Similarly, one should note that the performance of the Gibbs sampled matrix method for the other eight datasets is similar to that shown in Figure 4. The average A_{roc} values for the Gibbs sampled matrix, TEPITOPE and the ClustalW weight-matrix, respectively, are 0.755, 0.703 and 0.692 for the complete datasets and 0.690, 0.630 and 0.637 for the reduced datasets. One other striking observation from Figure 6 is the poor performance of the TEPITOPE method on the Southwood dataset. A simple calculation outlines a possible explanation for this poor performance. If one calculates the odds (frequency/background) values for the amino acid composition at the possible P1 positions in the Southwood dataset, one finds that the three amino acids with the highest odds ratios are F, W and Y. This stands in contrast to the finding in the other datasets, where no particular bias is found in the amino acids with the highest odds. The amino acid composition bias at the P1 position in the Southwood dataset originates from the selection bias in the prediction algorithm used to select the peptides for binding assay verification (Southwood *et al.*, 1998). In the TEPITOPE weight-matrix, the P1 position is modeled in a very crude manner, in that all non-hydrophobic amino acids have a value of -999 and the hydrophobic amino acids have a value of either 0 (F, W and Y) or -1 (I, L, M and V). In the Gibbs sampler matrix, this picture is more differentiated. Here, the difference in weight-matrix score between the common (I, L, M and V) and the rare amino acids (F, W and Y) is on an average 10. The importance of this distinction between the different amino acids allowed becomes clear if one sets the P1 weight-matrix values for F, Y and W of the TEPITOPE matrix to nine. Using the modified TEPITOPE matrix, the A_{roc} value is increased to 0.80. The average performance on the other datasets in the benchmark calculation is comparable with that of the original TEPITOPE matrix.

DISCUSSION

We have developed a refined and specialized Gibbs sampling method for detecting the binding motif for MHC classes I and II. For the method to handle situations when only very few data points are available and to limit any sequence redundancy in the training dataset, we implemented sequence weighting and pseudo-count correction schemes. The parameters in these schemes were optimized using sets of pre-aligned sequences known to bind to MHC class I. The best performance was found for a sequence-weighting scheme based on sequence clustering. Pseudo-count correction for low counts with pseudo-count frequencies estimated as described by Altschul *et al.* (1997) was found to improve the prediction accuracy substantially. Prior knowledge of important positions in the sequence motif was also included to guide the motif-search by allowing for differential weighting of high information positions in the motif. A position specific weighting scheme with a 3-fold increase in the weight of high information position in the motif as compared with the background was found to be optimal. The performance of predictors based on these principles had for most of the evaluation sets better predictive performance than what could be obtained using the HMMER and the SYFPEITHI prediction methods.

Using these parameters, we developed a method for detecting binding motifs in peptides known to bind to MHC II. These peptides are typically longer than the core motif, and correct alignment is key for obtaining good prediction performance. The optimal Gibbs sampler solution (the one with the highest information content) is not necessarily the optimal predictor, and we showed that including suboptimal solutions in an ensemble average increased the predictive performance of the method. We compared the performance of our Gibbs sampler with the TEPITOPE method and a ClustalW derived weight-matrix in a large-scale benchmark calculation using 10 datasets. In all cases, the ClustalW weight-matrix has a performance that is lower than that of the Gibbs sampled matrix. Only for one of the 10 datasets does the TEPITOPE weight-matrix have a predictive performance value that is significantly higher than the Gibbs sampled matrix. This was found to be due to a much higher cysteine content in the test set than in our training set. The TEPITOPE method arbitrarily assigns a zero weight to all cysteines. If we force the Gibbs sampler to treat cysteines as alanines, we obtain a higher performance than TEPITOPE for all test sets.

Prediction of class II MHC epitopes is a difficult task, and the prediction accuracy of the method described is far from perfect. At least two avenues exist where one can expect to achieve higher accuracy prediction methods. One avenue is the development of more sophisticated methods. Earlier, we have shown for MHC class I that a combination of many artificial neural networks with different types of sequence encoding leads to predictors of improved accuracy (Nielsen *et al.*, 2003). Using the Gibbs sampler as an alignment pre-processing as described by Brusic *et al.* (1998a), a similar approach might be beneficial for MHC class II predictions. A second avenue to improved prediction algorithms is a generation of relevant training data. For MHC class I, we have shown that the use of quantitative binding data as opposed to classification data leads to higher accuracy predictors (Buus *et al.*, 2003; Nielsen *et al.*, 2003). Furthermore, we have demonstrated that a guided iterative training process where new data points are selected from experimental binding assay verification by the method of query by committee (QBC) can in a highly cost and time-efficient manner lead to high accuracy prediction methods (Christensen *et al.*, 2003). Likewise, we believe that a similar approach can be applied to the MHC class II problem. The weight matrix obtained by the Gibbs sampler can generate first generation peptide predictions to be verified in binding affinity assays. Subsequently, the QBC method can guide the process of generating highly informative data that upon experimental verification effectively can provide high-quality prediction methods.

We have illustrated the use of the Gibbs sampler in the problem of identifying potential class I and class II MHC epitopes, but it should be valid for other applications where the sequence motif is weak and alignment is a crucial part of the motif identification.

ACKNOWLEDGEMENTS

This work was supported by the Danish National Research Foundation, the Danish MRC (grant 22-01-0272), the 5th Framework programme of the European Commission (grant QLRT-1999-00173) and the NIH (grant AI49213-02).

REFERENCES

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.

Brusic,V., Petrovsky,N., Zhang,G. and Bajic,V.B. (2002) Prediction of promiscuous peptides that bind HLA class I molecules. *Immunol. Cell Biol.*, **80**, 280–285.

Brusic,V., Rudy,G. and Harrison,L.C. (1994) Prediction of MHC binding peptides using artificial neural networks. In Stonier,R.J. and Yu,X.S. (eds), *Complex Systems: Mechanism of Adaptation*. IOS Press, Amsterdam, Holland, pp. 253–260.

Brusic,V., Rudy,G., Honeyman,G., Hammer,J. and Harrison,L. (1998a) Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network. *Bioinformatics*, **14**, 121–130.

Brusic,V., Rudy,G. and Harrison,L.C. (1998b) MHCPEP, a database of MHC-binding peptides: update 1997. *Nucleic Acids Res.*, **26**, 368–371.

Buus,S., Lauemøller,S.L., Worning,P., Kesmir,C., Frimurer,T., Corbet,S., Fomsgaard,A., Hilden,J., Holm,A. and Brunak,S. (2003) Sensitive quantitative predictions of peptide-MHC binding by a ‘Query by Committee’ artificial neural network approach. *Tissue Antigens*, **62**, 378–384.

Castellino,F., Zhong,G. and Germain,R.N. (1997) Antigen presentation by MHC class II molecules: invariant chain function, protein trafficking, and the molecular basis of diverse determinant capture. *Hum. Immunol.*, **54**, 159–169.

Chenna,R., Sugawara,H., Koike,T., Lopez,R., Gibson,T.J., Higgins,D.G. and Thompson,J.D. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.*, **31**, 3497–3500.

Christensen,J.K., Lamberth,K., Nielsen,M., Lundegaard,C., Worning,P., Lauemøller,S.L., Buus,S., Brunak,S. and Lund,O. (2003) Selecting informative data for developing peptide-MHC binding predictors using a “Query by Committee” approach. *Neural comput.*, **15**, 2931–2942.

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Geluk,A., van Meijgaarden,K.E., Schloot,N.C., Drijfhout,J.W., Ottenhoff,T.H. and Roep,B.O. (1998) HLA-DR binding analysis of peptides from islet antigens in IDDM. *Diabetes*, **47**, 1594–1600.

Hammer,J., Bono,E., Gallazzi,F., Belunis,C., Nagy,Z. and Sinigaglia,F. (1994) Precise prediction of major histocompatibility complex class II-peptide interaction based on peptide side chain scanning. *J. Exp. Med.*, **180**, 2353–2358.

Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci., USA*, **89**, 10915–10919.

Henikoff,S. and Henikoff,J.G. (1994) Position-based sequence weights. *J. Mol. Biol.*, **243**, 574–578.

Hobohm,U., Scharf,M., Schneider,R. and Sander,C. (1992) Selection of representative protein datasets. *Protein Sci.*, **1**, 409–417.

Kullback,S. and Leibler,R.A. (1951) On information and sufficiency. *Ann. Math. Stat.*, **22**, 76–86.

Lawrence,C.E., Altschul,S.F., Boguski,M.S., Liu,J.S., Neuwald,A.F. and Wootton,J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.

Marshall,K.W., Wilson,K.J., Liang,J., Woods,A., Zaller,D. and Rothbard,J.B. (1995) Prediction of peptide affinity to HLA DRB1*0401. *J. Immunol.*, **154**, 5927–5933.

Metropolis,N., Rosenbluth,A.W., Teller,A.H. and Teller,E. (1953) Equation of state calculation by fast computing machines. *J. Chem. Phys.*, **21**, 1087–1092.

Neuwald,A.F., Liu,J.S. and Lawrence,C.E. (1995) Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci.*, **4**, 1618–1632.

Nielsen,M., Lundegaard,C., Worning,P., Lauemoller,S.L., Lamberth,K., Buus,S., Brunak,S. and Lund,O. (2003) Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci.*, **12**, 1007–1017.

Parker,K.C., Bednarek,M.A. and Coligan,J.E. (1994) Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. *J. Immunol.*, **152**, 163–175.

Press,W.H., Flannery,B.P., Teukolsky,S.A. and Vetterling,W.T. (1993) *Numerical Recipies*, 2nd edn. Cambridge University Press, Cambridge, UK.

Rammensee,H., Bachmann,J., Emmerich,N.P., Bachor,O.A. and Stevanovic,S. (1999) SYFPEITHI: database for MHC ligands and peptide motifs. Immunogenetics, **50**, 213–219.

Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.

Serwold,T., Gonzalez,F., Kim,J., Jacob,R. and Shastri,N. (2002) ERAAP customizes peptides for MHC class I molecules in the endoplasmic reticulum. *Nature*, **419**, 480–483.

Sette,A. and Sidney,J. (1999) Nine major HLA class I supertypes account for the vast preponderance of HLA-A and -B polymorphism. *Immunogenetics*, **50**, 201–212.

Sette,A., Adorini,L., Colon,S.M., Buus,S. and Grey,H.M. (1989a) Capacity of intact proteins to bind to MHC class II molecules. *J. Immunol.*, **143**, 1265–1267.

Sette,A., Buus,S., Appella,E., Smith,J.A., Chesnut,R., Miles,C., Colon,S.M. and Grey,H.M. (1989b) Prediction of major histocompatibility complex binding regions of protein antigens by sequence pattern analysis. *Proc. Natl Acad. Sci., USA*, **86**, 3296–3300.

Southwood,S., Sidney,J., Kondo,A., del Guercio,M.F., Appella,E., Hoffman,S., Kubo,R.T., Chesnut,R.W., Grey,H.M. and Sette,A. (1998) Several common HLA-DR types share largely overlapping peptide binding repertoires. *J. Immunol*, **160**, 3363–3373.

Sturniolo,T., Bono,E., Ding,J., Raddrizzani,L., Tuereci,O., Sahin,U., Braxenthaler,M., Gallazzi,F., Protti,M.P., Sinigaglia,F. and Hammer,J., (1999) Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. *Nat. Biotechnol.*, **17**, 555–561.

Swets,J.A. (1988) Measuring the accuracy of diagnostic systems. *Science*, **240**, 1285–1293.

Sylvester-Hvid,C., Kristensen,N., Blicher,T., Ferre,H., Lauemoller,S.L., Wolf,X.A., Lamberth,K., Nissen,M.H., Pedersen,L.O. and Buus,S. (2002) Establishment of a quantitative ELISA capable of determining peptide–MHC class I interaction. *Tissue Antigens*, **59**, 251–258.

Thompson,W., Rouchka,E.C. and Lawrence,C.E. (2003) Gibbs recursive sampler: finding transcription factor binding sites. *Nucleic Acids Res.*, **31**, 3580–3585.

Yewdell,J., Anton,L.C., Bacik,I., Schubert,U., Snyder,H.L. and Bennink,J.R. (1999) Generating MHC class I ligands from viral gene products. *Immunol Rev.*, **172**, 97–108.