

An integrative approach to CTL epitope prediction: A combined algorithm integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions

Mette Voldby Larsen¹, Claus Lundegaard¹, Kasper Lamberth², Søren Buus², Søren Brunak¹, Ole Lund¹ and Morten Nielsen¹

¹ Center for Biological Sequence Analysis, BioCentrum-DTU, Technical University of Denmark, Lyngby, Denmark

² Institute for Medical Microbiology and Immunology, Panum Institute, Copenhagen, Denmark

Reverse immunogenetic approaches attempt to optimize the selection of candidate epitopes, and thus minimize the experimental effort needed to identify new epitopes. When predicting cytotoxic T cell epitopes, the main focus has been on the highly specific MHC class I binding event. Methods have also been developed for predicting the antigen-processing steps preceding MHC class I binding, including proteasomal cleavage and transporter associated with antigen processing (TAP) transport efficiency. Here, we use a dataset obtained from the SYFPEITHI database to show that a method integrating predictions of MHC class I binding affinity, TAP transport efficiency, and C-terminal proteasomal cleavage outperforms any of the individual methods. Using an independent evaluation dataset of HIV epitopes from the Los Alamos database, the validity of the integrated method is confirmed. The performance of the integrated method is found to be significantly higher than that of the two publicly available prediction methods BIMAS and SYFPEITHI. To identify 85% of the epitopes in the HIV dataset, 9% and 10% of all possible nonamers in the HIV proteins must be tested when using the BIMAS and SYFPEITHI methods, respectively, for the selection of candidate epitopes. This number is reduced to 7% when using the integrated method. In practical terms, this means that the experimental effort needed to identify an epitope in a hypothetical protein with 85% probability is reduced by 20–30% when using the integrated method.

The method is available at <http://www.cbs.dtu.dk/services/NetCTL>. Supplementary material is available at <http://www.cbs.dtu.dk/suppl/immunology/CTL.php>.

Received 29/10/04

Revised 27/4/05

Accepted 7/6/05

[DOI 10.1002/eji.200425811]

Key words:

CTL epitope · MHC class I binding · Proteasomal cleavage · TAP transport efficiency · Artificial neural networks

Introduction

For the cytotoxic T cells (CTL) of the immune system to discriminate between healthy cells and infected cells, all

nucleated cells present a selection of the peptides contained in their proteins on the cell surface in complex with MHC class I. However, only a small fraction of the peptides in a pathogen proteome are able to elicit a CTL response. This is mainly due to the selectivity in the antigen-processing steps preceding the CTL response. For each MHC class I allele, only 1 out of 2000 potential peptides will be immunodominant [1]. A prerequisite for the induction of a CTL response is the generation of peptides from their precursor polypeptides. The major cytosolic protease associated with the generation of antigenic peptides – in particular the C-terminal end of the peptides – is the proteasome [2–6]. After protea-

Correspondence: Morten Nielsen, Center for Biological Sequence Analysis, BioCentrum-DTU, Building 208, Technical University of Denmark, DK-2800 Lyngby, Denmark
Fax: +45-4593-1585
e-mail: mniel@cbs.dtu.dk

Abbreviations: ANN: artificial neural network · AUC: area under the ROC curve · TAP: transporter associated with antigen processing

somal cleavage the peptides may be trimmed at the N-terminal end by other peptidases in the cytosol [7]. The next step is the translocation of the peptides from the cytosol to the interior of the ER. This transport is facilitated by binding of the peptides to TAP. Once inside the ER, further N-terminal trimming of the peptides may occur [5, 8], as well as binding of some of the peptides to MHC class I. After binding, the MHC class I:peptide complex is transported to the surface of the cell, where it may be recognized by CTL. The most restrictive step involved in antigen presentation is the binding to MHC class I. It is estimated that only 1 out of 200 peptides will bind a given MHC class I allele with sufficient strength to elicit a CTL response [1]. However, it has previously been shown that the proteasomal cleavage and the TAP transport efficiency show also some degree of specificity [9, 10].

Reliable predictions of immunogenic peptides can minimize the experimental effort needed to identify new epitopes. Accordingly, many attempts have been made to predict the outcome of the steps involved in antigen presentation. A number of methods have been developed that very reliably predict the binding affinity of peptides to the different MHC class I alleles [11–14]. Likewise, methods have been developed that predict the efficiency with which peptides of arbitrary length will be transported by TAP [15, 16]. Several methods have also been developed that aim at predicting the proteasomal cleavage pattern of proteins. One such method is NetChop, which can be found in various versions, trained on different types of data [17]. In this work, we focus on NetChop C-term and NetChop 20S. NetChop C-term has been trained on natural MHC class I ligands, whereas NetChop 20S is trained on *in vitro* cleavage data. It has previously been shown that NetChop 20S is less accurate in predicting the C-terminal ends of naturally occurring MHC class I ligands than NetChop C-term 2.0 [17]. Similar results were obtained in a new implementation of NetChop (NetChop C-term 3.0 and NetChop 20S-3.0), where a superior performance was obtained using a novel network training strategy and sequence encoding scheme [18].

Combining HLA-A*0201 affinity predictions with predictions of TAP transport efficiency has previously been done by Peters *et al.* [16], and was shown to lead to improved identification of CTL epitopes. In the same work, they also combined HLA-A*0201 affinity predictions with predictions of C-terminal cleavages by NetChop 20S, and showed that this led to a less accurate identification of epitopes. The analysis of Peters *et al.* was, however, limited to a single MHC class I allele. Here, we extended the analysis to include epitopes for a large set of MHC class I supertypes [19, 20]. Further, we modeled the proteasomal cleavage event by novel and more

reliable prediction algorithms [18]. We generated a dataset containing 148 nonameric epitopes extracted from the SYFPEITHI database (<http://www.syfpeithi.de>, [21]). The majority of these peptides have successfully passed the steps involved in antigen presentation. We used this dataset to develop an improved method for CTL epitope identification by combining the prediction methods for MHC class I affinity, TAP transport efficiency, and C-terminal cleavage, and have demonstrated that the integrative approach has a predictive performance that is superior to predictions of MHC class I affinity alone using either artificial neural networks (ANN) [13], BIMAS [22], or SYFPEITHI [23]. Internal cleavage of peptides may result in the destruction of epitopes. However, we have found that prediction of such internal cleavage sites does not improve the predictability of epitopes. Finally, we have confirmed the validity of our integrated method on an independent dataset of HIV epitopes from the Los Alamos database. Here, we show that the performance of the integrated method is superior to the two publicly available methods BIMAS [22] and SYFPEITHI [23].

Results

We defined a combined prediction score for MHC class I affinity, TAP transport efficiency, and C-terminal proteasomal cleavage as a weighted sum of the three individual prediction scores. For MHC class I affinity, we used the rescaled prediction values, as described in Materials and methods. For TAP transport efficiency we used the method of Peters *et al.* [16], and for the proteasomal cleavage one of the four cleavage predictors described in Materials and methods. We used the SYF¹ dataset to estimate the set of weights in which the Arank and AUC values were optimal (see Materials and methods for a description of Arank and AUC values). The optimal combined method was found to have relative weights on TAP transport efficiency and C-terminal cleavage of 0.05 and 0.1, respectively. Note that these weights do not directly reflect the relative contribution of different prediction methods since the individual methods do not give comparable output values. Strong internal cleavage sites could destroy potential epitopes. We therefore tested if predictions of internal cleavage sites could contribute to the identification of epitopes when combined with predictions of MHC class I binding. However, we found that none of the internal sites could improve the ability to identify epitopes (data not shown). In Fig. 1, we show examples of ROC and rank curves for the SYF¹ dataset (see Materials and methods). The figure shows the performance curves for six different prediction scoring schemes: INT, NetMHC, TAP, NetChop3.0, NetCh-

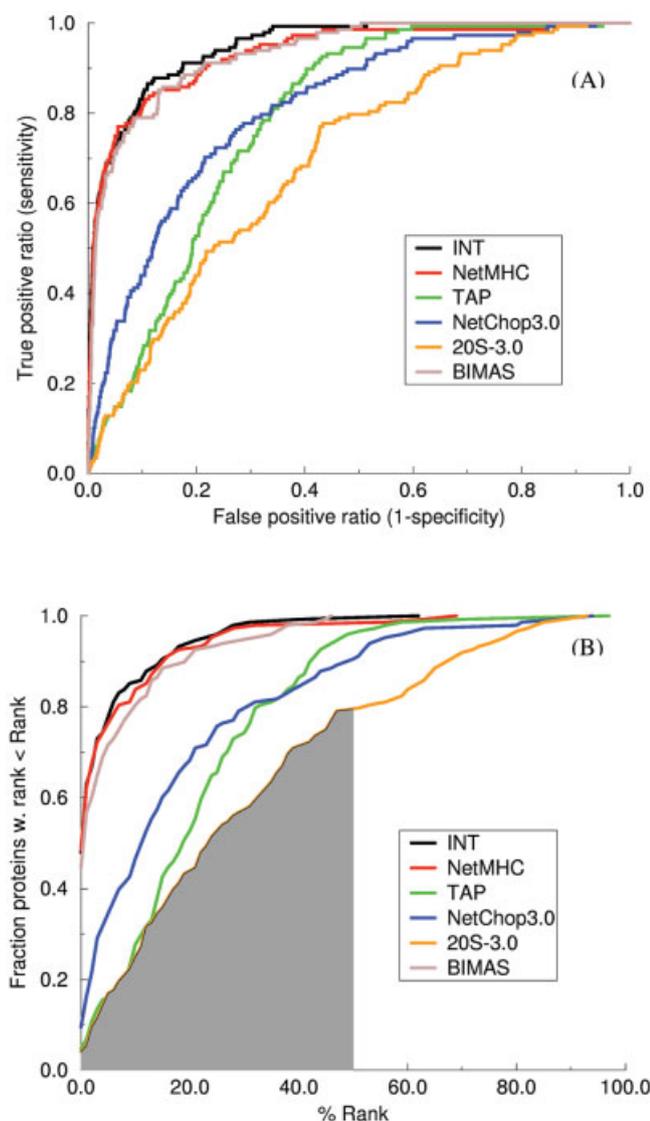


Fig. 1. ROC and Rank performance curves for different prediction methods. (A) The ROC curves. AUC is the area under the ROC curve. (B) Rank curves. Arank is the area under the rank curve (as an example, the area under the 20S-3.0 curve is highlighted) as described in Materials and methods. Predictions are made on the SYF¹ dataset. The different prediction methods are; INT: Optimal integrated method with relative weight on TAP transport efficiency and C-terminal cleavage of 0.05 and 0.1, respectively. NetMHC: MHC class I affinity by ANN, TAP: TAP transport efficiency. NetChop3.0: C-terminal cleavage of the NetChop C-term 3.0 proteasomal predictor. 20S-3.0: C-terminal cleavage of the NetChop 20S-3.0 predictor, and BIMAS: MHC class I binding predictor [22].

op20S-3.0, and BIMAS. Here, the INT method is the integrated method with relative weight on TAP and NetChop3.0 of 0.05 and 0.1, respectively, NetMHC represents ANN-based MHC class I affinity predictions, NetChop3.0 and NetChop20S-3.0 are the C-terminal cleavage predictions of NetChop C-term 3.0 and NetChop 20S-3.0, respectively, and BIMAS is the publicly available method for MHC peptide binding

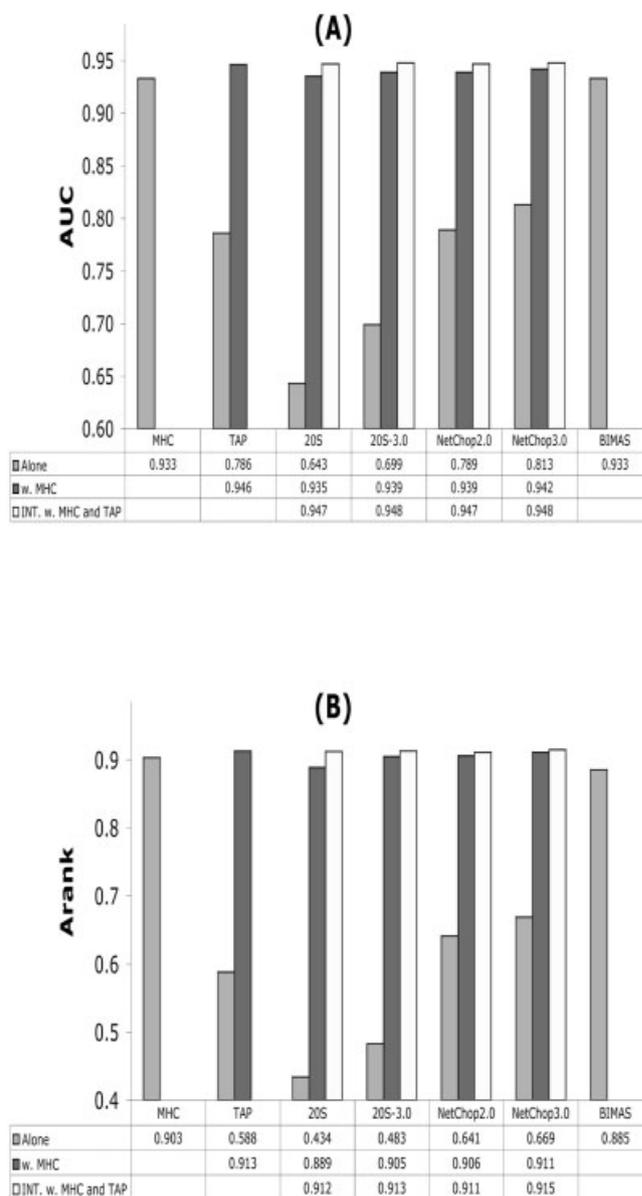


Fig. 2. Predictive performance for different prediction methods. (A) AUC is the area under the ROC curve, and (B) Arank is the area under the rank curve as described in Materials and methods. Predictions are made on the SYF¹ dataset. The figure shows the performance measures for each method on its own, the optimal performance in combination with MHC class I affinity predictions, and the optimal performance in combination with TAP transport efficiency and MHC class I affinity predictions. The different prediction methods are; NetMHC: MHC class I affinity by ANN, TAP: TAP transport efficiency, 20S: C-terminal cleavage of the NetChop 20S-2.0 predictor, 20S-3.0: C-terminal cleavage of the NetChop 20S-3.0 predictor, NetChop2.0: C-terminal cleavage of the NetChop C-term 2.0 proteasomal predictor, NetChop3.0: C-terminal cleavage of the NetChop C-term 3.0 proteasomal predictor. BIMAS is a publicly available MHC class I binding prediction method [22].

predictions [22]. Fig. 2 gives the details of the performance measures for the different methods and their combinations when evaluated on the SYF¹ dataset.

The ROC curves in Fig. 1A highlights a problematic aspect of using the AUC performance measure when dealing with highly unbalanced datasets. The AUC values for the TAP and NetChop C-term 3.0 prediction methods are close to identical (0.79 and 0.81, see Fig. 2A). However, looking at the curves for each method, it is clear that the NetChop C-term 3.0 method provides the most useful predictions. The region of the ROC curve where the TAP predictor performs best falls in a highly non-relevant region of the specificity. The two curves cross at a false-positive ratio of 0.35. This value corresponds to 35% false-positive predictions, and having an improved prediction method only in this specificity range is clearly irrelevant. For the Arank measure this problem is not present since we here explicitly focus on the high rank region only.

The results shown in Figs 1 and 2 demonstrate that the method integrating predictions of MHC class I affinity, TAP transport efficiency, and proteasomal cleavage has the highest performance in terms of both the AUC and Arank values. The individual method with the poorest performance is that of NetChop 20S, followed by NetChop 20S-3.0, TAP, the NetChop C-term 2.0 and NetChop C-term 3.0 methods, and MHC class I affinity (BIMAS and NetMHC). Performing a bootstrap experiment to determine which method has the highest predictive performance using the Arank measure, we found that NetMHC performs marginally better than the BIMAS method ($p=0.06$). The integrated INT method, on the other hand, has a performance on both the AUC and Arank measures that is significantly higher than that of both the BIMAS and NetMHC prediction methods ($p<0.01$ in all comparisons).

It is striking to observe that even though the different NetChop predictors individually have very different predictive performance, they all achieve similar predictive performance when combined with MHC class I affinity predictions. We see, for instance, only a marginally higher performance of the method integrating predictions from the best proteasomal cleavage prediction method NetChop C-term 3.0 relative to a combined method integrating predictions from the poorest proteasomal cleavage prediction method NetChop 20S. The NetChop C-term 3.0 predictor, which has been trained on epitope data, has been criticized for predicting a combination of MHC class I affinity, TAP transport efficiency, and proteasomal cleavage rather than just proteasomal cleavage [16]. Here, we find that the NetChop 20S-3.0 and TAP predictors can be combined in a constructive manner with a predictive performance significantly higher than that of the individual predictors (AUC=0.805, Arank=0.615). This is not the case for the NetChop C-term 3.0 predictor. Here, the combination with TAP only leads to a minor and insignificant improvement in the predictive perfor-

mance (AUC=0.813, Arank=0.670). However, when combined with NetMHC and TAP, both NetChop 20S-3.0 and NetChop C-term 3.0 achieve similar predictive performance, suggesting that NetChop C-term 3.0 has a proteasomal cleavage signal with a quality that is comparable to that of NetChop 20S-3.0, which has been trained on *in vitro* cleavage data.

In Fig. 3, we compare the predictive performance of the method of NetMHC, BIMAS, SYFPEITHI, and that of the integrated method when evaluated on the SYF² dataset. Using the bootstrap analysis on the Arank performance measure, we found that NetMHC performs significantly better than SYFPEITHI ($p<0.001$), and marginally better than the BIMAS method ($p=0.05$). The integrated INT method has a performance on both the AUC and Arank measures that is significantly higher than that of the SYFPEITHI, BIMAS, and NetMHC prediction methods ($p<0.01$ in all comparisons).

A direct measure of the performance gain when comparing the integrated method to that of predicted MHC affinity predictions alone is the rank value needed to identify 85% of the epitopes in a dataset. For the SYF² dataset, this rank value is 23%, 12%, and 12% when using the prediction methods of SYFPEITHI, BIMAS, and NetMHC, respectively. For the integrated method, this rank value is reduced to 8%. Evaluating the different prediction methods on the larger SYF¹ dataset, we found similar results. Here, the 85% coverage rank value for the BIMAS, NetMHC and integrated method is 12%, 12%, and 8%, respectively. We define the term reliability of a prediction method as the probability of identifying an epitope in a given protein within a certain top percentage of the peptides. If a protein contains 300 nonameric peptides, 24, 36, 36 and 64 peptides must be

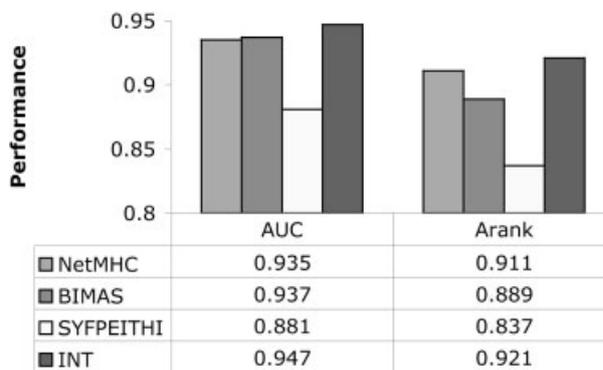


Fig. 3. AUC and Arank performance measures for the different prediction methods evaluated on the SYF² dataset. The methods included in the figure are; NetMHC: MHC class I affinity by ANN, BIMAS and SYFPEITHI: Two publicly available methods predicting MHC class I binding [22, 23], and INT: The integrated method combining predictions of MHC class I binding, TAP transport efficiency, and C-terminal proteasomal cleavage.

selected for experimental verification to reach 85% reliability, if using the integrated, NetMHC, BIMAS, and SYFPEITHI prediction methods, respectively. Compared to the BIMAS method, the integrated method will thus, on average, lead to a reduction in the experimental effort by more than 30%. If the experimental effort is limited to the peptides predicted to be in the top 5%, the reliability values for the integrated, NetMHC, BIMAS, and SYFPEITHI methods, respectively, are 78%, 76%, 72%, and 69%.

Next, we repeated the analysis using the benchmark data from the independent HIV dataset. The overall AUC and Arank values of the analysis are visualized in Fig. 4. The results confirm the findings from the SYF datasets that the performance of the integrated method is superior to that of the individual methods. We found that the integrated method performs better than the BIMAS and SYFPEITHI methods ($p < 0.025$). The number of peptides needed to be tested to reach 85% reliability is 7%, 10%, and 9% for the integrated, NetMHC, and BIMAS predictions, respectively, when analyzing the complete HIV¹ dataset with 69 epitopes. Looking at the reduced HIV² dataset and including the SYFPEITHI prediction method, the corresponding numbers are 7%, 9%, and 10% for the integrated, BIMAS, and SYFPEITHI method, respectively. Finally, we find that the reliability values when looking only at the top-scoring 5% predictions are 77% (77%), 73% (73%), and 72% for the integrated, BIMAS, and

SYFPEITHI methods, respectively. Here, the values in parenthesis refer to the complete HIV¹ dataset. These numbers thus support the results found when using the SYF datasets.

Discussion

In this report, we have used an integrative approach to improve CTL epitope identification. We have integrated predictions of MHC class I binding affinity, TAP transport efficiency, and C-terminal proteasomal cleavage, and demonstrated in a large-scale benchmark that the integrated method has a predictive performance significantly higher than any of the individual methods, as well as the publicly available BIMAS and SYFPEITHI prediction methods.

Other groups have previously combined different prediction methods: Hakenberg *et al.* [24] developed a bioinformatical tool for prediction of CTL epitopes by combining predictions of proteasomal cleavage and MHC class I affinity. On a very small dataset of only five epitopes from HIV Nef, Kesmir *et al.* [17] showed that combining predictions of proteasomal cleavage with measured TAP and MHC class I binding affinity correlates well with the observed number of MHC class I ligands presented on the cell. In another study, Peters *et al.* [16] improved identification of epitopes by combining predictions of binding affinities to the HLA-A*0201 allele with predictions of TAP transport efficiency. They also combined HLA-A*0201 affinity predictions with predictions of C-terminal cleavages by NetChop 20S, and showed that this lead to a less accurate identification of epitopes.

In the present work, we have extended the analysis of Peters *et al.* to include epitopes from ten different MHC class I supertypes spanning the large variation in MHC class I specificity. Further, we have modeled the proteasomal cleavage event by novel and more reliable prediction algorithms [18]. Including a broad set of MHC class I specificities in the analysis allows us to: (1) draw more general and well-founded conclusions about how to integrate the different steps in the class I pathway in the most optimal manner, and (2) derive a prediction method that is broadly applicable for the identification of CTL epitopes.

In designing the optimal prediction method, we tested several version of the NetChop cleavage predictor. We found that the two NetChop methods trained on *in vitro* digestion data on their own have the poorest performance, followed by the two NetChop methods trained on epitope data. However, when combining cleavage predictions with affinity to MHC class I and TAP transport efficiency, all combinations achieve close to identical performance. Concern has previously been

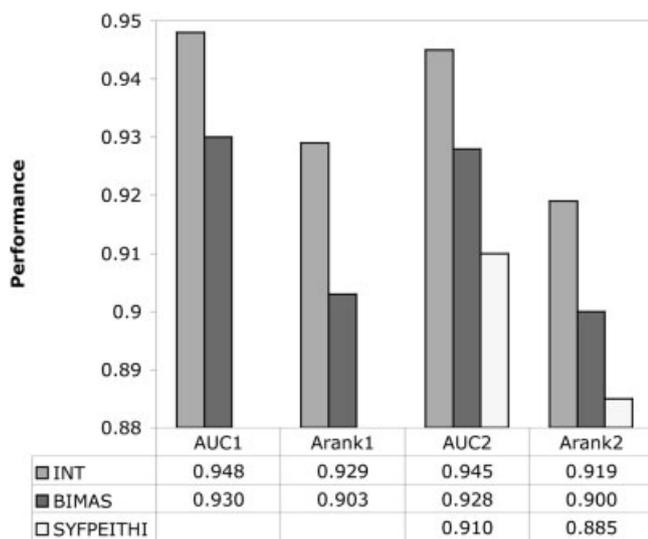


Fig. 4. Predictive performance for the different prediction methods. Predictions are made on the two HIV datasets. AUC1 and Arank1 refer to the complete HIV¹ dataset, AUC2 and Arank2 to the reduced HIV² dataset. The methods included in the figure are: INT: The integrated method integrating predictions of MHC class I binding TAP transport efficiency, and C-terminal proteasomal cleavage, BIMAS and SYFPEITHI: Two publicly available methods predicting MHC class I binding [22, 23].

raised that the NetChop method, which has been trained on natural MHC class I ligand data does not only predict proteasomal cleavage, but rather a combination of cleavage, TAP transport efficiency, and affinity to the “average” MHC class I allele [16]. Here, we demonstrate that when predicting CTL epitopes, the NetChop method trained on epitope data outperforms the methods trained on *in vitro* degradation data. However, in combination with MHC class I affinity and TAP transport efficiency predictions both epitope and *in vitro* digest trained methods show similar performance. Two conclusions can be drawn from this: (1) the high performance of the NetChop method trained on epitope data does not come from more accurate predictions of the proteasomal cleavage event, but rather from indirect integration of TAP transport efficiency and MHC class I affinity, and (2) the proteasomal cleavage predicting element of the NetChop method trained on epitope data has a quality that is comparable to that of the method trained on *in vitro* data. These observations leave, however, promise for future improvements to CTL epitope predictions, since it should be possible to improve proteasomal predictions by developing a method describing the specificity of the immunoproteasome and integrate it with NetChop 20S-3.0, which predicts the constitutive proteasome cleavage specificity.

We have validated the performance of the integrated method on a set of known HIV epitopes derived from the Los Alamos database. A direct implication of the improved predictive performance of the integrated method is a significant gain in sensitivity at 85% reliability level. Using the integrated method, we thus find that the experimental effort needed to reach 85% reliability is significantly reduced as compared to predictions based on MHC class I affinity alone (using NetMHC, BIMAS, or SYFPEITHI prediction methods). We believe that this improved identification of peptides capable of eliciting a CTL response will be useful in reverse immunogenetic approaches, and hence in the process of rational vaccine design.

Materials and methods

Datasets

SYF dataset

On February 5, 2004, 779 nonameric peptides present in the SYFPEITHI database (<http://www.syfpeithi.de>) [21] and classified as either ligand for MHC class I (HLA-A or HLA-B) or T cell epitope were extracted. To obtain the largest possible set of data, we included peptides classified as T cell epitopes in the analyses, even though these peptides are not for certain naturally processed. This we did at the risk of potentially

including a small set of misclassified data. Since the non-natural pathway most likely does not share any of the specificities of the class I pathway (except MHC class I binding) such data will lower the significance of our analyses. In the SYFPEITHI database, it is stated that, “as far as T cell epitopes are concerned, only those have been selected which are likely to be naturally processed”, and the potential number of false classifications should hence be small. For every peptide, the source protein was subsequently found in the SwissProt database. If more than one protein was the possible origin of a given peptide, a protein was chosen according to the following criteria: It had to be either a human protein or a protein from a human pathogen. If there were still more possible proteins, the correct one was found by tracking the source of the peptide in the SYFPEITHI database. The resulting dataset contained 663 peptides with corresponding source proteins. In Table 1, the peptides have been grouped according to MHC class I allele, and further grouped into one of the 12 MHC class I supertypes [20]. Since we only have access to high accuracy prediction methods for the 12 MHC class I supertypes, and their alleles, peptides binding to alleles like HLA-A23, HLA-A2902, HLA-B14, and HLA-B1516 that are not classified as belonging to any of the supertypes were excluded from the study. Within each supertype duplicate peptides were removed, leaving 567 peptides in the dataset. After removing the peptides that have been used to train the MHC class I affinity-predictors or NetChop C-term 2.0/3.0 prediction methods, 148 peptides remained in the dataset. These peptides are referred to as the epitopes of the SYF¹ dataset. SYFPEITHI [21] does not have predictors for any alleles belonging to the supertypes B58 and B62, and to be able to compare the performance against SYFPEITHI predictions, a subset (SYF²) was constructed where peptides restricted to alleles assigned to these supertypes were excluded. This dataset contains 140 peptides.

HIV dataset

On May 12, 2004, 1005 nonameric peptides present in of the HIV Immunology CTL database of the Los Alamos HIV Database (www.hiv.lanl.gov) were collected. For the HIV epitope data, we have no direct handle on whether a peptide has been naturally processed or not. A non-naturally processed epitope will in the analysis appear as a false positive, and lower the significance of our results. To achieve a large dataset we, however, chose to include all peptides in the analysis. Peptides known only to bind non-human MHC class I or binding an unknown MHC class I allele were removed from the dataset. As with the peptides in the SYF dataset, the peptides were grouped according to MHC class I allele, and further grouped into 1 of the 12 MHC class I supertypes. Peptides binding to alleles that are not classified as belonging to any of the supertypes were excluded from the study. Subsequently, all the peptides that have been used to train the MHC class I affinity-predictors or NetChop C-term 2.0/3.0 were removed from the set. The amino acid sequences of the 9 SwissProt entries from taxon Human immunodeficiency virus type 1 (HXB2 isolate) (HIV-1) were then collected, and only the 69 peptides contained in one of these proteins were included in the final dataset. These 69 peptides are referred to as the epitopes of

Table 1. Summary of alleles selected to represent the different supertypes^{a)}

Supertype	Alleles included in supertype	No of 9mers ^{b)}	No of 9mers not used for training ^{c)}	NetMHC	BIMAS	SYF
A1	A01, A3001, A3002, A3003, A3004	26	1	A0101	A1	A*01
A2	A0201, A0204, A0205, A0206, A0207, A0214, A0217, A6901	172	39	A0201	A0201	A*0201
A3	A03, A0301, A1101, A3101, A3303, A6601, A6801	56	25	A1101	A3	A*03
A24	A24, A2402, A2403	38	15	A2403	A24	A*2402
A26	A2601, A2602, A2603	5	–	–	–	–
B7	B07, B0702, B35, B3501, B3503, B5101, B5102, B5103, B53, B5301, B5501, B5502, B5601	87	34	B0702	B7	B*0702
B8	B08, B0801, B0802	22	5	B0801	B8	B*08
B27	B1518, B27, B2701, B2702, B2703, B2704, B2705, B2706, B7301	62	18	B2705	B2705	B*2705
B39	B1509, B1510, B3801, B3901, B3909	29	–	–	–	–
B44	B3701, B40, B40012, B4006, B44, B4402, B4403	20	3	B4001	B40	B*4402
B58	B1513, B1517, B5701, B5702, B5801, B5802	25	4	B5801	B5801	–
B62	B1501, B1502, B1503, B1508, B1512, B4601, B5201	25	4	B1501	B62	–
TOTAL	67 (75)	567	148			

^{a)} The 12 supertypes and the alleles are classified as belonging to each supertype [20].

^{b)} The number of unique nonamers in each group for which it was possible to locate a source protein in the SwissProt database.

^{c)} The number of nonamers not included in training of NetMHC and NetChop C-term 2.0/3.0 predictors. The last three columns summarize which NetMHC, BIMAS [22] and SYFPEITHI [23] methods are used to represent each supertype.

dataset HIV¹. To be able to compare the performance against SYFPEITHI predictions, a subset (HIV²) was constructed where peptides restricted to alleles assigned to the B58 and the B62 supertypes were excluded. The HIV² dataset contained 62 peptides. For both the SYF and HIV datasets all nonameric peptides contained in the protein sequences from which the epitopes originated, except those annotated as epitopes in either the complete SYFPEITHI or Los Alamos HIV databases, were taken as negative peptides and will be referred to as non-epitopes. When using this definition of epitopes/non-epitopes one has to take into account that some nonamers will falsely be classified as non-epitopes because the SYFPEITHI and Los Alamos HIV databases are incomplete. Since the MHC class I molecules are very specific, binding only a highly limited repertoire of peptides, this misclassified proportion will, however, be very small. A given MHC class I molecule has a specificity of ~1% [1]. In a protein of 100 amino acids, one expects to have 1 binding and approximately 99 non-binding peptides. The potential number of false classifications is hence orders of magnitude smaller than the actual number of negatives. All datasets are available as complementary material at <http://www.cbs.dtu.dk/suppl/immunology/CTL.php>.

Predicting proteasomal cleavage patterns

The four prediction methods described below were used individually to assign a predicted cleavage value to the residues of the proteins in the SYF and HIV dataset. A low predicted cleavage value corresponds to a low probability of proteasomal cleavage, whereas a high value corresponds to a high probability of cleavage.

The C-term 2.0 and 20S networks of the NetChop 2.0 prediction server (www.cbs.dtu.dk/services/NetChop): This is a standard artificial feed-forward neural network (ANN) with one hidden layer of units. C-term 2.0 is trained on publicly available human MHC class I ligands. The C-terminal amino acids of the ligands were classified as cleavage sites, whereas sites within the ligands were labeled non-cleavage sites. 20S is trained on constitutive proteasome *in vitro* digests of yeast enolase and bovine-casein. For details on the dataset construction and ANN training see [17].

The NetChop C-term 3.0 and NetChop 20S-3.0 prediction servers (www.cbs.dtu.dk/services/NetChop-3.0): These ANNs have been trained on the same data as the NetChop C-term 2.0 and NetChop 20S networks, using an optimized training strategy and combinations of several ANNs each trained with different types of sequence encodings. In a large-scale benchmark calculation these new networks were shown to have a predictive performance superior to that of both NetChop and other public methods [18].

Predicting TAP transport efficiency

The TAP transport efficiency prediction method is based on the matrix described in Peters *et al.* [16]. Predicted TAP transport efficiency of peptides with arbitrary length is calculated by scoring only the C terminus and the three N-terminal residues. The contribution of the N-terminal residues to the final score is down-weighted by a factor of 0.2 in comparison with the contribution of the C terminus. The TAP transport efficiency score for a given nonamer is given as the average of the values for the nonamer and its decameric precursor. When using the method, a low predicted value corresponds to a low TAP transport efficiency, whereas a high predicted value corresponds to a high TAP transport efficiency. Notice that in the article by Peters *et al.* [16] the reverse situation applies, but we have multiplied all numbers in the matrix by a factor of -1 to facilitate the later combination of the TAP transport efficiency with the proteasomal cleavage pattern and MHC class I binding affinity.

MHC class I affinity predictions

For the A1, A2, A3, A24, B7, B8, B27, B44, B58, and B62 supertype the affinity predictors are based on ANN [12, 13]. This prediction method is available at <http://cbs.dtu.dk/services/NetMHC>. Each supertype is represented by an ANN trained on nonameric peptides with known binding affinity to a given MHC class I allele. Table 1 summarizes which alleles are selected to represent the different superotypes. Predictors for the A26 and B39 superotypes were not available, and peptides binding alleles of either of these two superotypes were excluded from the study. Each peptide is assigned a value between 0 and 1, where 0 corresponds to low MHC class I affinity and 1 to high affinity [13]. When combining the predictions of MHC class I affinity, TAP transport efficiency, and proteasomal cleavage, the MHC class I affinities were rescaled to make the prediction values comparable between MHC class I superotypes. Sturniolo *et al.* [25] have outlined a simple approach to perform such a rescaling: For each predictor, the MHC class I affinities for 500 000 random peptides were predicted and the 1% fractals found. The rescaling is then performed by a simple division of predicted MHC class I affinity by the 1% fractal for the corresponding supertype. We shall refer to the ANN-based MHC class I affinity prediction method as NetMHC. To validate the predictive performance of the developed method, its performance was compared to that of the BIMAS [22] and SYFPEITHI [23] prediction methods. Table 1 summarizes which prediction matrices were used for BIMAS and SYFPEITHI for the different superotypes.

Performance measures

We applied a series of performance measures and statistical tests to evaluate the predictive performance of the different methods used in this study. When combining MHC class I affinity predictions with predicted TAP transport efficiency and C-terminal proteasomal cleavage, we applied two non-parametric performance measures. One measure is the conventional AUC value (the area under the ROC curve) [26]. In this measure, all overlapping nonameric peptides in

the dataset are sorted according to the prediction score. The epitopes define the positive set, whereas the negative set is made up from the non-epitopes. In a typical calculation for the SYF dataset for instance, the positive set contained 148 peptides, and the negative set more than 92 000 peptides. The ROC curve is plotted from the sensitivity and 1-specificity values calculated by varying the cut-off value (separating the predicted positive from the predicted negative) from high to low. The area under this curve gives the AUC value. The AUC value is 0.5 for a random prediction method and 1.0 for a perfect method. Even though commonly used, the AUC measure is not easy to interpret intuitively. In the AUC measure, all predictions are sorted in one pool. In situations where the data consist of several proteins of very different length, each with only one positive example, the longer proteins will clearly contribute in a biased way to the AUC measure. Here, we therefore designed a second performance measure with a clearer and more intuitive interpretation. This measure is a rank measure. For each protein in the benchmark, we sorted all nonameric peptides based on the prediction score. The rank value for the protein was calculated as the percentage of non-epitope peptides with a score higher than that of the corresponding epitope. From these rank values, we constructed a rank curve showing the accumulative fraction of proteins with a rank value below a certain value. From the rank-curve, one can extract information on how large a fraction of the proteins will have their epitope within top 5% of the predictions for instance. Finally, we defined a single performance measure (Arank) as the area under the rank-curve integrated from rank zero up to rank 50%. A perfect prediction method will have all the epitopes at the top of the sorted list with a rank of 0, and thus an Arank value of 1.0, whereas a poor method will have the epitopes placed randomly in the sorted list and hence have an Arank value of 0.25. The Arank value is calculated for rank value up to 50% only, to focus on the top rank part of the curve. A method that improves the rank position of an epitope within the top 50% is clearly more relevant than a method improving the rank position in the bottom end of the rank curve. Examples of ROC and rank-curves are shown in Fig. 1. For both the AUC and Arank performance measure, we were aware that some nonamers will falsely be classified as non-epitopes because the SYFPEITHI and Los Alamos HIV databases are incomplete.

The Bootstrap method used to test the significance of a performance gain

We applied the Bootstrap method [27] to test the significance of the difference in predictive performance between two prediction methods. In a Bootstrap experiment, we generated a series of evaluation set replications by randomly drawing n data points with replacement from the original dataset, where n is the size of the original dataset. For each dataset replica, we evaluated the predictive performance (AUC or Arank) of the two methods, and the Bootstrap hypothesis test p value for the hypothesis that method M1 performs better than method M2 is estimated from the simple ratio $\#(M1 < M2)/N$, where $\#(M1 < M2)$ is the number of experiments where method M2 outperformed method M1, and N the number of Bootstrap

replica. A *p* value of 0.05 will indicate that method M1 is performing significantly better than method M2.

Acknowledgements: This project was in part funded by Genomes2Vaccines (STREP), FP6, contract no.: LSHB-CT-2003–503231, and NIH Contract # HHSN266200400083C.

References

- 1 Yewdell, J. W. and Bennink, J. R., Immunodominance in major histocompatibility complex class I-restricted T lymphocyte responses. *Annu. Rev. Immunol.* 1999. **17**: 51–88.
- 2 Craiu, A., Akopian, T., Goldberg, A. and Rock, K. L., Two distinct proteolytic processes in the generation of a major histocompatibility complex class I presented peptide. *Proc. Natl. Acad. Sci. USA* 1997. **94**: 10850–10855.
- 3 Stoltze, L., Dick, T. P., Deeg, M., Pommerl, B., Rammensee, H. G. and Schild, H., Generation of the vesicular stomatitis virus nucleoprotein cytotoxic T lymphocyte epitope requires proteasome-dependent and -independent proteolytic activities. *Eur. J. Immunol.* 1998. **28**: 4029–4036.
- 4 Paz, P., Brouwenstijn, N., Perry, R. and Shastri, N., Discrete proteolytic intermediates in the MHC class I antigen processing pathway and MHC I-dependent peptide trimming in the ER. *Immunity* 1999. **11**: 241–251.
- 5 Mo, X. Y., Cascio, P., Lemerise, K., Goldberg, A. L. and Rock, K., Distinct proteolytic processes generate the C and N termini of MHC class I-binding peptides. *J. Immunol.* 1999. **163**: 5851–5859.
- 6 Altuvia, Y. and Margalit, H., Sequence signals for generation of antigenic peptides by the proteasome: implications for proteasomal cleavage mechanism. *J. Mol. Biol.* 2000. **295**: 879–890.
- 7 Levy, F., Burri, L., Morel, S., Peitrequin, A. L., Levy, N., Bachi, A., Hellman, U. et al., The final N-terminal trimming of a subamino-terminal proline-containing HLA class I-restricted antigenic peptide in the cytosol is mediated by two peptidases. *J. Immunol.* 2002. **169**: 4161–4171.
- 8 Stoltze, L., Schirle, M., Schwarz, G., Schroter, C., Thompson, M. W., Hersh, L. B., Kalbacher, H. et al., Two new proteases in the MHC class I processing pathway. *Nat. Immunol.* 2000. **1**: 413–418.
- 9 Kesmir, C., Nussbaum, A. K., Schild, H., Detours, V. and Brunak, S., Prediction of proteasome cleavage motifs by neural networks. *Protein. Eng.* 2002. **15**: 287–296.
- 10 Endert, v. P. M., Riganelli, D., Greco, G., Fleischhauer, K., Sidney, J., Sette, A. and Bach, J. F., The peptide-binding motif for the human transporter associated with antigen processing. *J. Exp. Med.* 1995. **182**: 1883–1895.
- 11 Brusic, V., Rudy, G. and Harrison, L. C., 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 1994, pp 253–260.
- 12 Buus, S., Laumoller, S. L., Worning, P., Kesmir, C., Frimurer, T., Corbet, S., Fomsgaard, A. et al., Sensitive quantitative predictions of peptide-MHC binding by a 'Query by Committee' artificial neural network approach. *Tissue Antigens* 2003. **62**: 378–384.
- 13 Nielsen, M., Lundegaard, C., Worning, P., Laumoller, S. L., Lamberth, K., Buus, S., Brunak, S. and Lund, O., Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci.* 2003. **12**: 1007–1017.
- 14 Nielsen, M., Lundegaard, C., Worning, P., Hvid, C. S., Lamberth, K., Buus, S., Brunak, S. and Lund, O., Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach. *Bioinformatics* 2004. **20**: 1388–1397.
- 15 Brusic, V., van Endert, P., Zeleznikow, J., Daniel, S., Hammer, J. and Petrovsky, N., A neural network model approach to the study of human TAP transporter. *In Silico Biol* 1999. **1**: 109–121.
- 16 Peters, B., Bulik, S., Tampe, R., Endert, P. M. V. and Holzhutter, H. G., Identifying MHC class I epitopes by predicting the TAP transport efficiency of epitope precursors. *J. Immunol.* 2003. **171**: 1741–1749.
- 17 Kesmir, C., Nussbaum, A. K., Schild, H., Detours, V. and Brunak, S., Prediction of proteasome cleavage motifs by neural networks. *Protein Eng.* 2002. **15**: 287–296.
- 18 Nielsen, M., Lundegaard, C., Lund, O. and Kesmir, C., The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage. *Immunogenetics* 2005. **57**: 33–41.
- 19 Sette, A. and Sidney, J., HLA supertypes and supermotifs: a functional perspective on HLA polymorphism. *Curr. Opin. Immunol.* 1998. **10**: 478–482.
- 20 Lund, O., Nielsen, M., Kesmir, C., Petersen, A. G., Lundegaard, C., Worning, P., Sylvester-Hvid, C. et al., Definition of supertypes for HLA molecules using clustering of specificity matrices. *Immunogenetics* 2004. **55**: 797–810.
- 21 Rammensee, H., Bachmann, J., Emmerich, N. P., Bachor, O. A. and Stevanovic, S., SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics* 1999. **50**: 213–219.
- 22 Parker, K. C., Bednarek, M. A. and Coligan, J. E., Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. *J. Immunol.* 1994. **152**: 163–175.
- 23 Rammensee, H. G., Bachmann, J. and Stevanovic, S., *MHC ligands and Peptide Motifs*. Chapman & Hall, New York: 1997.
- 24 Hakenberg, J., Nussbaum, A. K., Schild, H., Rammensee, H. G., Kuttler, C., Holzhutter, H. G., Kloetzel, P. M. et al., MAPPP: MHC class I antigenic peptide processing prediction. *Appl. Bioinformatics* 2003. **2**: 155–158.
- 25 Sturniolo, T., Bono, E., Ding, J., Radrizzani, L., Tuereci, O., Sahin, U., Braxenthaler, M. et al., Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. *Nat. Biotechnol.* 1999. **17**: 555–561.
- 26 Swets, J. A., Measuring the accuracy of diagnostic systems. *Science* 1988. **240**: 1285–1293.
- 27 Press, W. H., Flannery, B. P., Teukolsky, S. A. and Vetterling, W. T., *Numerical Recipes in C: The Art of Scientific Computing*, 2nd Edn. Cambridge University Press, Cambridge 1992