Immunogenetics (2012) 64:177–186
DOI 10.1007/s00251-011-0579-8

ORIGINAL PAPER

NetMHCcons: a consensus method for the major histocompatibility complex class I predictions

Edita Karosiene · Claus Lundegaard · Ole Lund · Morten Nielsen

Received: 2 July 2011 / Accepted: 28 September 2011 / Published online: 20 October 2011

Abstract A key role in cell-mediated immunity is dedicated to the major histocompatibility complex (MHC) molecules that bind peptides for presentation on the cell surface. Several *in silico* methods capable of predicting peptide binding to MHC class I have been developed. The accuracy of these methods depends on the data available characterizing the binding specificity of the MHC molecules. It has, moreover, been demonstrated that consensus methods defined as combinations of two or more different methods led to improved prediction accuracy. This plethora of methods makes it very difficult for the non-expert user to choose the most suitable method for predicting binding to a given MHC molecule. In this study, we have therefore made an in-depth analysis of combinations of three state-of-the-art MHC–peptide binding prediction methods (*NetMHC*, *NetMHCpan* and PickPocket). We demonstrate that a simple combination of *NetMHC* and *NetMHCpan* gives the highest performance when the allele in question is included in the training and is characterized by at least 50 data points with at least ten binders. Otherwise, *NetMHCpan* is the best predictor. When an allele has not been characterized, the performance depends on the distance to the training data. *NetMHCpan* has the highest performance when close neighbours are present in the training set, while the combination of *NetMHCpan* and *PickPocket* outperforms either of the two methods for alleles with more remote neighbours. The final method, NetMHCcons, is publicly available at www.cbs.dtu.dk/services/NetMHCcons, and allows the user in an automatic manner to obtain the most accurate predictions for any given MHC molecule.

Keywords MHC class I · T cell epitope · MHC binding specificity · Peptide–MHC binding · Consensus methods · Artificial neural network

Introduction

Major histocompatibility complex (MHC) molecules play a key role in cell-meditated immunity binding antigenic peptides and presenting them for recognition by the immune system on the cell surface. Through antigen processing, proteins produced within a cell are degraded into short peptides, usually of 8–11 residues in length that may then be loaded on MHC-I molecules and presented on the cell surface. In this way, cytotoxic T lymphocytes are capable of recognizing the infected cells and triggering an immune response. Thousands of different allelic versions of MHC molecules exist (Robinson et al. 2001), making complete experimental characterization of peptide–MHC interactions highly cost-intensive. A number of *in silico* prediction methods for peptide–MHC binding have therefore been successfully developed during the last decade (for a review, see, e.g., Lundegaard et al. 2010). It has been demonstrated that the predictive performance of MHC peptide binding prediction methods depends strongly on both the number of peptides and the number of actual binders available for training (Yu et al. 2002; Zhang et al. 2009a, b). For pan-specific methods, the performance has moreover been demonstrated to dependent strongly on the

Electronic supplementary material The online version of this article (doi:10.1007/s00251-011-0579-8) contains supplementary material, which is available to authorized users.

E. Karosiene (✉) · C. Lundegaard · O. Lund · M. Nielsen
Center for Biological Sequence Analysis,
Department of Systems Biology,
Technical University of Denmark,
Building 208, Kemitorvet,
Lyngby 2800, Denmark
e-mail: edita@cbs.dtu.dk

amino acid sequence distance to the nearest allelic neighbour in the data used to train the method (Hoof et al. 2009; Zhang et al. 2009a). Moreover have several benchmark studies shown that consensus methods defined as a simple average of two or more different methods can lead to improved prediction accuracy (Moutaftsi et al. 2006; Wang et al. 2008; Zhang et al. 2009a, b). This means that one method or sets of methods may perform well for one given MHC molecule while performing poorly for others. Even though several benchmarks have been carried out to compare MHC binding methods and rank them based on their prediction accuracy (Lin et al. 2008; Peters et al. 2006; Zhang et al. 2009a, b), it remains a highly non-trivial task for the end-user to select the best suitable method for a given MHC molecule.

The objective of this study was to address this problem and define a method that for any given MHC molecule in an automatic manner defines an optimal combination of a series of prediction methods, allowing the non-expert end-user to obtain accurate binding predictions. Three state-of-the-art methods *NetMHC*, *NetMHCpan* and *PickPocket* were included in this study. *NetMHC* is an artificial neural network-based (ANN) allele-specific method, capable of predicting binding only to the molecules on which it has been trained (Lundegaard et al. 2008; Nielsen et al. 2003). The two other methods are pan-specific meaning that they are able to predict peptide binding also to MHC molecules for which limited or no experimental peptide binding data is available. NetMHCpan, is ANN-based (Hoof et al. 2009; Nielsen et al. 2007), and the *PickPocket* method is matrix-based and relies on receptor-pocket similarities between MHC molecules (Zhang et al. 2009a). The choice of methods to be analyzed was made based on previous benchmark studies. *NetMHC* and *NetMHCpan* methods have in several large-scale benchmark studies been demonstrated to be among the best publically available predictors (Lundegaard et al. 2010; Peters et al. 2006; Zhang et al. 2009b). Even though the *PickPocket* method has not in any benchmark studies been shown to provide a superior performance, it has been demonstrated that the method for alleles with no close neighbours can improve binding affinity predictions when combined with *NetMHCpan* (Zhang et al. 2009a). All the methods were benchmarked using a large and diverse set of quantitative peptide binding affinity measurements, covering more than 100 MHC class I alleles.

It is apparent that not all methods can be applied to predict binding to any chosen MHC molecule. For example, the *NetMHC* method is available only if the allele in question is also part of the training set used to develop the method. On the other hand, the pan-specific *NetMHCpan* and *PickPocket* methods are capable of predicting binding to any MHC molecule with known protein sequence. The development of the consensus method was guided by simplicity and robustness. This means that the combination of two or more methods was only included into definition of the final method if it demonstrated a significantly improved performance compared to the individual methods within the analysed conditions. In the paper, we first benchmark each method individually and evaluate their performance under different settings. Next, given these results, the consensus method is defined in an allele-specific manner as a combination of one or more prediction methods, and finally is the consensus method, NetMHCcons, validated against an independent data set.

Materials and methods

Data set

The benchmark data set consists of quantitative nonameric peptide–MHC class I binding data with a submission date prior to September 2009 retrieved from the IEDB (Vita et al. 2010) and an in-house MHC–peptide binding database. In total, it consists of 101,728 unique peptide–MHC class I interactions covering 101 alleles: 34 HLA-A, 35 HLA-B, one HLA-C, one HLA-E, 11 chimpanzee (Patr), 12 rhesus macaque (Mamu), one gorilla (Gogo), and six mouse alleles. Table S1 contains a detailed description of the benchmark data set. All peptide binding measurements were obtained as IC_{50}/EC_{50} values and for this study were log-transformed to fall in the range between 0 and 1 using the relation $1-\log(IC_{50}\text{nM})/\log(50{,}000)$ (Nielsen et al. 2003).

Analyzed methods and conditions

For our analysis, we used in-house versions of *NetMHC*, *NetMHCpan* and *PickPocket* trained and evaluated on the MHC class I benchmark data set. Having both allele-specific (*NetMHC*) and pan-specific (*NetMHCpan* and PickPocket) methods in the benchmark resulted into two analyzed conditions: (1) when allele in question is part of the training set; (2) when allele in question is not part of the training set.

When an allele for which the binding should be predicted was not part of the training data, the analysis reduced to include only the two pan-specific methods. In all other cases, the analysis included all three methods. In order to obtain a reliable performance when evaluating the methods, we constructed a reduced data set consisting of 78 alleles, for which at least 50 data points were available and at least ten of them were binding peptides (i.e., having an affinity stronger than 500 nM). This reduced data set is presented in Table S2.

Evaluation strategies

When training ANNs, it is critical to define a strategy to avoid overfitting. Conventionally, this is done using a test set to stop the network training when the performance on the test set is optimal. This is a highly CPU-intensive procedure since the evaluation must be made using nested cross-validation. For 5-fold nested cross-validation for instance, the data is split into five subsets. In each round, one subset is employed as evaluation set and is not included into training process. The remaining four subsets are used in the inner cross-validation loop where four networks are constructed each using three sets to train the network and one set used to stop the training to avoid overfitting. The binding predictions of the peptide in the evaluation set are next calculated as a simple average of the four networks in each cross-validation ensemble. Another faster strategy for cross-validation is to use the test set as evaluation data. In this setting, the test set is used both to stop the network training in order to avoid overfitting, and to evaluate the predictive performance. This cross-validation approach has an inherent potential of overestimating the predictive performance.

To evaluate to what degree the use of the faster training strategy led to an overestimation of the predictive performance, the two different evaluation strategies were compared for the *NetMHC* and *NetMHCpan* methods in terms of Pearson's correlation coefficients (PCC). As the *PickPocket* method has no stopping procedure, this method was not included in this comparison.

Our analysis showed that the difference between evaluation on independent data sets compared to evaluation on test sets during cross-validation is significant for the *NetMHC* method when the data set is smaller than 1,000 data points ($p=0.02$), and not significant for the sets with 1,000 or more data points ($p=0.15$). For the *NetMHCpan* method, which always has a large evaluation set, no difference in performance was observed between the two evaluation strategies ($p=0.19$) (see Fig. S1). As a result of this and in order to reduce the computational efforts, we have for the further analysis chosen to use independent set for *NetMHC* evaluations and make the *NetMHCpan* evaluations on the test sets during cross-validation, which is a much faster approach.

In order to evaluate the predictive performance of *NetMHCpan* and *PickPocket* for alleles, which are not part of the training data, a leave-one-out (LOO) approach was used, meaning that the data for the allele in question was excluded from the training set. One important characteristics of the benchmark data set is that many peptides have been tested for binding to multiple MHC molecules. Given this nature of the peptide data set, it is essential to design the LOO training strategy so that not only data for the specific allele in question is removed from the training, but also peptides common between the evaluation and training sets. In doing this, we assure that neither the MHC molecule nor the peptides are present in the training and evaluation sets at the same time. In order to avoid reducing the training set too much in this strict LOO setup, the evaluation set was split into three subsets and the performance for each subset was evaluated. Setting the number of evaluation subset partitions to three was based on a compromise between increase in calculation time and the accuracy of performance estimation. A small number of subset partitions would be computationally fast but would lead to relative large reductions in the size of the training data, and likewise would a large number of subset partitions be computationally costly but lead to only a minor reduction in the size of the training data. A small evaluation of the performance as a function of the number of evaluation subset was carried out for the HLA-A*02:01 allele. This is the allele in the data set characterized by the largest number of peptide measurements, and hence is the example where the peptide overlap to other molecules should be the highest. This evaluation demonstrated that only limited gain in performance was achieved for subset divisions larger than three (data not shown), therefore leading us to use three subset through the benchmark evaluation.

Calculation of pseudo-distance to the nearest neighbour

Pseudo-distance between two alleles was calculated from the pseudo-sequences of MHC molecules as described in (Nielsen et al. 2007). The nearest neighbour for a specific allele is defined as the molecule from the neighbour reference which includes MHC molecules with more than 50 data points and more than ten binders (Hoof et al. 2009), with the smallest pseudo-distance to this allele.

Defining the consensus method

A consensus method is defined in terms of combinations of two or more different individual method. Here, we use a simple average of the raw log-transformed prediction scores from each method to define the consensus method. Combined methods are represented using a plus sign "+" in this study. For each allele, the performance of each prediction method and their possible combinations were given as PCC between the log-transformed predicted and measured binding affinities.

Validation of the consensus method

An independent evaluation set consisting of data from the IEDB (Vita et al. 2010) and an in-house MHC–peptide

binding database with a submission date after September 2009 was constructed. This validation data set had no overlap with the training set. In this way, we ensured that the final consensus method was not trained and evaluated using the same data points. In order to obtain reliable evaluations, the only alleles characterized by at least 10 data points and at least two binders were included. The validation data set included 14,923 peptide–MHC binding data and covered 62 alleles (see Table S3). Part of these alleles (46) were included in the training data set, hence allowing a validation of the consensus method in two conditions: (1) for the alleles in question being part of the training data and (2) for the novel alleles not described in the training data.

Statistical analysis

In this study, the evaluation of significance of the observed differences between the results was performed using one-tailed paired *t*-test with a significance level of 0.05. If a very low *p* value was obtained during analysis, it is then stated as "$p<0.0001$" and not by exact value.

Results

The objective of this study was to define a strategy that for any given MHC molecule defines an optimal combination of a series of prediction methods, allowing the non-expert end-user in an automated manner to obtain accurate binding predictions for any given MHC molecule. The "Results" section falls into three subsections. First, we illustrate the end-user problem of identifying which method to use for binding prediction for a given MHC molecule, next we analyse in a large-scale benchmark how a simple yet powerful setting can be defined leading to a consensus method that consistently outperform all single methods included in the benchmark, and finally the consensus method is validated on an independent data set of MHC peptide binding measurements not included in the method development.

Performance variations of different methods

The motivation to perform this study was based on earlier observations that different methods give different prediction results in different conditions. To illustrate this, we compared how two ANN-based methods (*NetMHC* and *NetMHCpan*) that were trained on identical data would handle a given prediction task. A protein sequence was submitted to the *NetMHC* and *NetMHCpan* methods, trained on the benchmark data set, and the methods were asked to predict binding to the HLA-B*38:01 molecule, which was defined by 136 peptide-MHC binding measurements of which only three were binders within the training data set. In the left panel of Fig. 1, the output of this analysis is represented as a scatter plot between the prediction values of *NetMHC* and *NetMHCpan*. It is apparent from the figure that the correlation between the prediction scores obtained by the two methods is low — the PCC is 0.569 — and that the difference is in particular large in the high binding tail of the two methods. In order to investigate the disagreement between these two methods in a more systematic manner, we obtained the predictions of both methods for all MHC molecules included in the training data. The right panel of Fig. 1 demonstrates how the correlation between the predictions by *NetMHC* and *NetMHCpan* methods depends on the size of the training set. For MHC molecules that are defined by few data points and have small number of actual binders, the correlation coefficient between *NetMHC* and *NetMHCpan* methods is very small. The difference between predictions of the two methods is diminished when the number of peptides and binders in the training set is increased.

Defining the consensus method

Allele in question is part of the training data

When an allele is part of the training data, all three methods and their combinations can be used to define the consensus method. Each method was evaluated using cross-validation on the benchmark data set. Figure 2 shows accumulative number of instances where each method achieved the highest predictive performance as a function of the number of data points and the number of binding peptides, respectively, characterizing the different MHC molecules. The figure clearly demonstrates that some methods achieve the highest performance more often than others and, thus, are more important for defining the optimal consensus method. Only one combination, *NetMHC* + *NetMHCpan*, consistently improved prediction accuracy. This combination has an increased accuracy for alleles characterized by a larger number of peptides and significantly outperforms both the *NetMHCpan* and *NetMHC* methods ($p=0.005$) for the set of alleles characterized by at least 500 data points. All in all, the *NetMHC* + *NetMHCpan* combination gives a superior prediction performance for most of the alleles from the benchmark data set. As can be seen in Fig. 2, *NetMHC* + *NetMHCpan* has the highest performance 60 times out of 92 (65.2%), excluding ties. The second best method is *NetMHCpan*, which achieves the highest prediction accuracy for the alleles with a little training data set and all in all gives the best scores for 18 alleles (19.6%). A similar tendency is observed when comparing the accumulative number of times any given method is winning as a function of the number of binding peptides within the training set (Fig. 2, right panel). It is striking to observe that both *NetMHC* and *PickPocket*

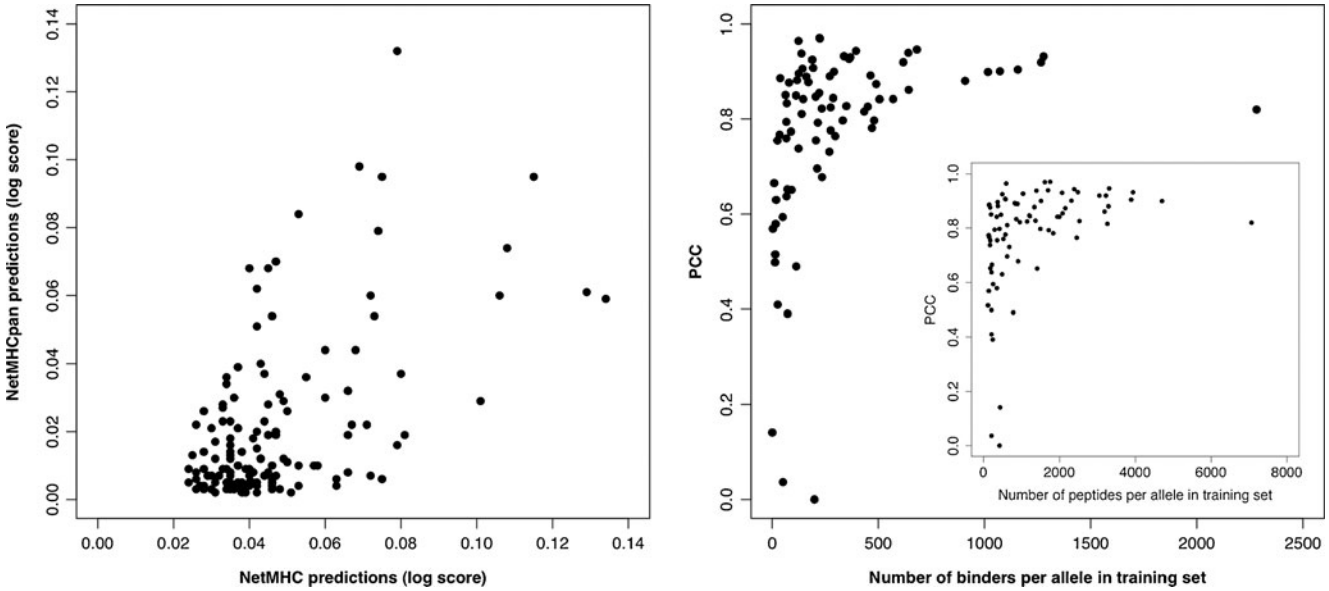

Fig. 1 Overview of agreement and disagreement between *NetMHC* and *NetMHCpan* methods. The *left panel* shows binding predictions to HLA-B*38:01 allele by *NetMHC* and *NetMHCpan* methods for peptides from the same chosen protein. Log-transformed prediction scores by each method are plotted. The *right panel* demonstrates the dependency of the Pearson's correlation coefficient between the two methods as a function of number of peptides (the inner plot) and number of binders available per allele in the training set

rarely perform best as single methods. Only when combined with *NetMHCpan* does *NetMHC* contribute to the overall performance and adding *PickPocket* seems to have a direct negative effect on the prediction accuracy.

The results displayed in Fig. 2 thus suggest *NetMHCpan* as the optimal method for alleles characterized by few peptide measurements, and a consensus method defined by *NetMHC* and *NetMHCpan* as optimal otherwise. To further investigate the precise setting for when to apply each of these methods, we compared the performance of *NetMHC*, *NetMHCpan* and *NetMHC* + *NetMHCpan* as a function of number of peptides (N_p) and number of binders (N_b), respectively, per allele in training set. The result of this analysis is given in Fig. 3. The analysis demonstrates that for alleles characterized by a small number of data points ($N_p<50$), the allele-specific *NetMHC* method performs poorly. In this case, the pan-specific *NetMHCpan* method clearly achieves the highest performance and significantly outperforms *NetMHC* ($p=0.02$). Considering the performance dependency on the number of binders per allele, one can notice that combination of the two methods always outperforms its components. The prediction accuracy of

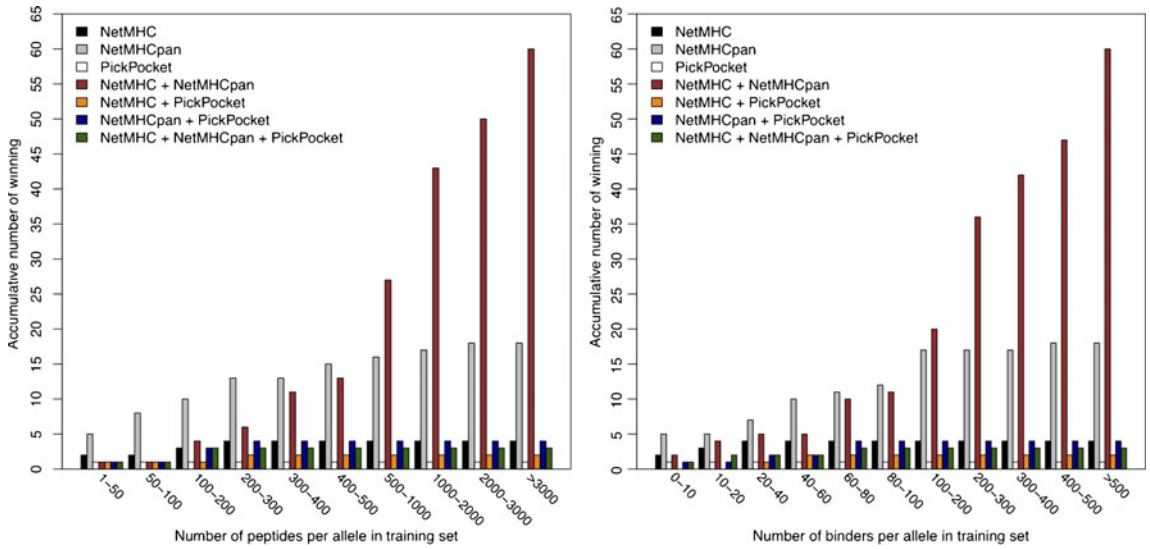

Fig. 2 Accumulative number of winning for each method included in the analysis depending on the number of peptides (*left*) and number of binders (*right*) per allele in training set

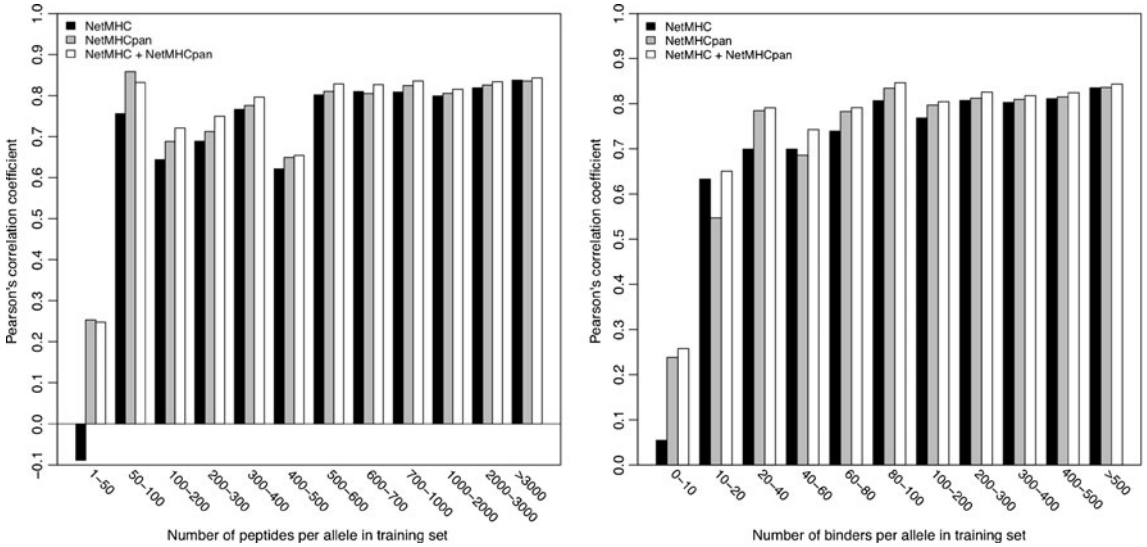

Fig. 3 The average predictive performance of the alleles in the benchmark data set as a function of number of peptides (*left*) and number of binders (*right*) per allele in the training data

the *NetMHC* method for MHC alleles characterized by few binders ($N_b<10$) is, however, very low, and the difference between *NetMHCpan* and the consensus method defined as *NetMHC* + *NetMHCpan* is for these alleles statistically insignificant. *NetMHCpan* also achieves the highest performance within the next bin ($50\leq N_p<100$). However, we cannot access the significance of the difference between the different methods in this case as we have only three alleles within the bin. Based on these observations and having in mind that we choose a single method where it is not significantly different from the combined approach, we defined the consensus method for the condition of the allele in question being part of the training data as follows:

$$\text{NetMHCcons} = \begin{cases} \text{NetMHCpan} & \text{for } N_p < 50 \text{ and } N_b < 10 \\ \text{NetMHC} + \text{NetMHCpan} & \text{otherwise} \end{cases}$$

Detailed results of the analysis of methods when allele in question is part of the training data are given in Table S4.

Allele in question is not part of the training data

When the MHC allele for which we wish to predict peptide binding is not part of the training data set, only the pan-specific *NetMHCpan* and *PickPocket* methods can be employed. It has been shown earlier that the predictive performance of pan-specific methods depends on the allelic environment. For example, *NetMHCpan* was demonstrated to perform well for the alleles with well-characterized neighbourhood (Hoof et al. 2009), and *PickPocket* was shown to give a good prediction accuracy for MHC molecules for which the similarity to characterized alleles was low (Zhang et al. 2009a). To investigate the performance of NetMHCpan, *PickPocket* and their combination, we conducted an LOO evaluation on the benchmark data set as described in "Materials and methods." The results are illustrated in Fig. 4 as the performance dependency on the distance to the nearest neighbour as measured in terms of the MHC pseudo-sequence similarity (detailed results of this analysis are presented in Table S5). The left panel of the figure demonstrates that a large fraction of the alleles from our benchmark data set have close nearest neighbours. Most of these alleles are human HLA-A and HLA-B alleles, whereas chimpanzee (Patr), macaque (Mamu) and mouse alleles tend to have more distant neighbours. It is apparent that the performance of both methods depends strongly on the distance from MHC molecule in question to the nearest molecule in the training set. Regression analysis for each method demonstrated, that the performance is decreased significantly with increasing distance for both methods ($p<0.0001$).

The right panel of the figure gives the average predictive performance of the different methods as a function of the distance to the nearest neighbour within the training data. The figure demonstrates the high performance of *NetMHCpan* in prediction of binding to MHC molecules with close neighbours. This method gives the highest PCC values of all methods when the distance (D) is lower than 0.1 and

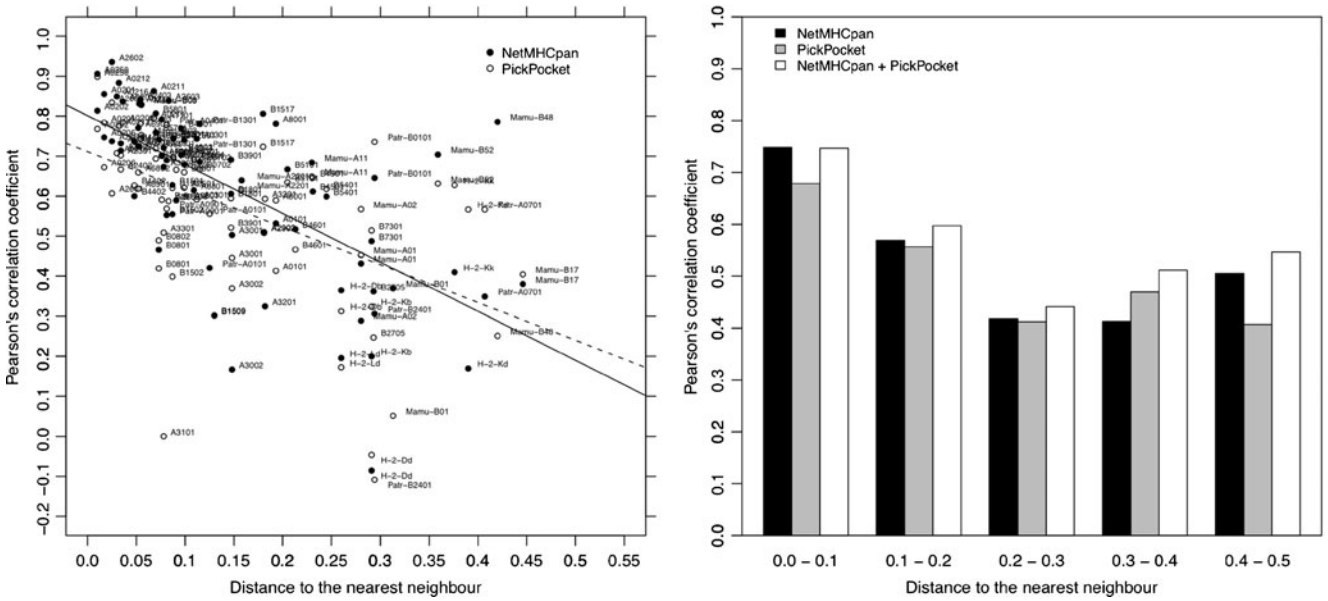

Fig. 4 Predictive performance of the alleles from the benchmark data set as a function of distance to the nearest neighbour. The *left panel* shows performance for each allele of the *NetMHCpan* and *PickPocket* methods. The *solid line* represents the least square fit for the *NetMHCpan* data, and the *dotted line* gives the least square fit for the *PickPocket* data. A full size of this graph is available in Fig. S2. The *right panel* demonstrates the average performance dependency on the distance to the nearest neighbour. The performance for each allele was calculated using leave-one-out approach as described in "Materials and methods." Distance to the nearest neighbour was calculated using MHC pseudo-sequences as described by Nielsen et al. (2007)

achieves the highest performance for 23 out of 40 MHC molecules, while *NetMHCpan* + *PickPocket* wins only 15 times within this bin. The difference between *NetMHCpan* and *NetMHCpan* + *PickPocket* was not statistically significant. If the distance to the nearest neighbour is larger than 0.1, the combination of the *NetMHCpan* and *PickPocket* methods significantly outperforms both *NetMHCpan* (p= 0.019) and *PickPocket* (p=0.003) methods. Based on these observations and our decision to use the simpler method where the significant difference is not observed, we define the optimal method to predict peptide binding to MHC molecules not included in the training set as follows:

$$\text{NetMHCcons} = \begin{cases} \text{NetMHCpan} & \text{for } D < 0.1 \\ \text{NetMHCpan} + \text{PickPocket} & \text{for } D \geq 0.1 \end{cases}$$

Based on the results obtained above, we can now define the final consensus method. We define a reference set of alleles that are characterized by at least 50 data points and at least ten binders. Based on this reference set, the *NetMHCcons* method can be defined as

$$\text{NetMHCcons} = \begin{cases} \text{NetMHC} + \text{NetMHCpan} & \text{for } D = 0 \\ \text{NetMHCpan} & \text{for } 0 < D < 0.1 \\ \text{NetMHCpan} + \text{PickPocket} & \text{for } D \geq 0.1 \end{cases}$$

where D refers to the distance between the query allele and its nearest neighbour in the reference allele set. Note that having the distance equal to 0 (D=0) means that the alleles in question is part of the training set.

Validation of the final consensus method

The consensus method for peptide binding to MHC was next benchmarked on an independent evaluation data set (see "Materials and methods"). In order to compare the results with the methods composing *NetMHCcons*, we obtained predictions of each method separately and compared the results for the subsets of alleles depending on how each method was involved in the final consensus method. This resulted into three different comparisons of the average PCC values: (1) for the alleles that were part of the training set, the results of *NetMHCcons* were compared with the results obtained by *NetMHCpan* and *NetMHC* methods (41 allele); (2) *NetMHCcons* was compared with *NetMHCpan* for all the alleles from the validation set (62 alleles); (3) the comparison of the consensus method with *NetMHCpan* and *PickPocket* was done using the alleles that were not included in the training data set and had a distance of 0.1 or larger to the training reference set (17 alleles).

A summary of the validation results is given in Fig. 5 (details are given in Table S6). The performance of *NetMHCcons* on the alleles that were part of the training set was found to be significantly higher than both *NetMHC* (p<0.0001) and *NetMHCpan* (p=0.01). Comparing *NetMHCcons* and *NetMHCpan* performances using all the alleles, significant

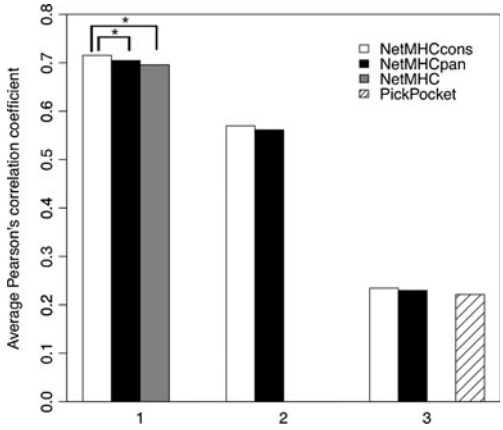

Fig. 5 Validation results of the *NetMHCcons* method. The plot shows three groups of comparisons, from the left: (1) *NetMHCcons*, *NetMHCpan* and *NetMHC* for the alleles common between training and validation sets; (2) *NetMHCcons* and *NetMHCpan* for all alleles in the validation set; (3) *NetMHCcons*, *NetMHCpan* and *PickPocket* for alleles not included in the training data and having $D \geq 0.1$. Significant difference between any two methods is indicated by *stars* and was calculated using paired one-tailed *t*-test

difference between performance values was not observed, but the consensus method has the highest performance of the two. The set of alleles that have a distance of 0.1 or more to the training data compose too small set to obtain significant p values; however, the average values show that the consensus method also here has a higher performance than both the *NetMHCpan* and *PickPocket* methods.

The *NetMHCcons* method is implemented as a web server and is available online at: http://www.cbs.dtu.dk/services/NetMHCcons. The method provides affinity predictions for any peptide of length 8–11 amino acids to any given MHC class I molecule of known protein sequence. Two submission types are handled — a list of peptides or a protein in FASTA format. The server provides a possibility for the user to choose MHC molecule in question from a list of alleles or alternatively upload the MHC protein sequence of interest. The method is also implemented as SOAP based Web Service available at: http://www.cbs.dtu.dk/ws/NetMHCcons/.

Discussion

In this study, we performed a detailed analysis of several state-of-the-art methods with a purpose of developing a consensus method that consistently provides the most accurate predictions for any given MHC molecule. To the best of our knowledge, this study analyzing and combining several different methods in an allele-specific manner is the first of its kind. Having involved allele-specific (*NetMHC*) and pan-specific (*NetMHCpan* and *PickPocket*) methods, two different conditions were analyzed in our study. First of all, if the given MHC allele had earlier been characterized, then all three methods and their combinations were analyzed. Here, we found that the prediction accuracy of the allele-specific *NetMHC* method depended strongly on the number of data available characterizing the given allele and demonstrated that for MHC molecules that are poorly characterized, the *NetMHCpan* method is the best predictor. On the other hand, increasing number of data points and binders available for the MHC molecule in question, the *NetMHC* method becomes important and the combination of this method with *NetMHCpan* provides the most accurate predictions. These conclusions are in agreement with an earlier report (Zhang et al. 2009b).

The vast majority of MHC molecules remain uncharacterized in terms of their binding specificity. For this reason, several pan-specific methods have been developed (Jacob and Vert 2008; Jojic et al. 2006; Nielsen et al. 2007; Zhang et al. 2009a). Moreover, several publications have demonstrated the importance of describing the subtle differences in binding specificity between MHC molecules in order to understand cellular immune responses of a given host to an infection (Erup Larsen et al. 2011; Hoof et al. 2010; Rapin et al. 2010; Stranzl et al. 2010). In our analysis, we considered two of the pan-specific methods *NetMHCpan* and *PickPocket*, both being able to produce high accuracy predictions for MHC molecules with limited or no binding data available. These methods were benchmarked under the conditions when the allele in question was not part of the training data set employing a LOO approach. We demonstrated that the performance of both methods reduces with increased distance to the nearest MHC molecule with characterized binding specificity. This is in agreement with previous studies (Zhang et al. 2009a). In our study, we additionally investigated how the performance of both methods and their combination depended on the distance to the closest characterized MHC molecule. At small distances, *NetMHCpan* demonstrated a superior performance, which was not maintained when the distance increased and at larger distances the contribution of the *PickPocket* method was demonstrated to be important when combined with the NetMHCpan. This is in accordance with the work by Zhang et al. (2009a). A consensus method defined as combination of *NetMHCpan* and *PickPocket* was hence shown to perform with the highest accuracy for MHC molecules with a large distance to MHC molecules with characterized binding specificity.

The final *NetMHCcons* method was validated using a diverse independent evaluation set. It was demonstrated that *NetMHCcons* achieved the highest performance compared with each separate method included in this analysis. This is, to our knowledge, the first consensus method defined as combination of three different methods, which involve both allele-specific and pan-specific approaches. Our analysis demonstrated how several methods could be combined into one capable of producing the most accurate predictions for any given allele. Such a method is of high importance to the non-

expert user allowing in an automated manner to obtain accurate predictions of binding to any MHC class I molecule of interest and also suggests that a similar approach might be employed to improve the accuracy of MHC class II predictors.

Several other high performing methods are publicly available for MHC class I predictions including the Average Relative Binding (ARB) matrix method (Bui et al. 2005) and the stabilized matrix method (SMM) (Peters and Sette 2005), both available as part of the IEDB tools for MHC class I binding prediction. None of these methods have been included in this study. In order to define a consensus method, a large independent evaluation data set is needed for obtaining a reliable performance estimate of the different methods and finding their optimal combination. The fact that the evaluation data must be large and independent makes it troublesome to include publicly available method in the benchmark analysis. If we define a large benchmark data, large parts of the data will most likely have been included in the training of the different methods and the evaluation will be erroneous due to over-fitting. If we limit ourselves to recently published data that most likely has not been included in the training of the different method, the evaluation data set become too small to allow for a robust method evaluation. Only by retraining the methods on the large data set can we maintain a large and independent evaluation data set allowing for a robust and unbiased evaluation of the different methods included in the benchmark. Including non-in-house methods would require expert knowledge of each method and hence must be carried out as a collaborative effort between the authors of the different method and is beyond the scope of this paper. We might suggest that such an effort should be carried out in the future along the lines of previous benchmark studies (Wang et al. 2010; Peters et al. 2006).

In conclusion, we have defined a method, *NetMHCcons*, in terms of the *NetMHCpan* method and its combinations with *NetMHC* and *PickPocket* based on conditions defining the MHC molecule in question. The method is implemented as a web server allowing the user in an automatic manner to obtain optimal predictions for any MHC class I molecule of interest.

Acknowledgements This work was supported by two NIH (National Institute of Health) grants (contract no. HHSN272200900045C, and contract no. HHSNN26600400006C).

References

Bui HH, Sidney J, Peters B, Sathiamurthy M, Sinichi A, Purton KA, Mothe BR, Chisari FV, Watkins DI, Sette A (2005) Automated generation and evaluation of specific MHC binding predictive tools: ARB matrix applications. Immunogenetics 57(5):304–314. doi:10.1007/s00251-005-0798-y

Erup Larsen M, Kloverpris H, Stryhn A, Koofhethile CK, Sims S, Ndung'u T, Goulder P, Buus S, Nielsen M (2011) HLArestrictor—a tool for patient-specific predictions of HLA restriction elements and optimal epitopes within peptides. Immunogenetics 63(1):43–55. doi:10.1007/s00251-010-0493-5

Hoof I, Perez CL, Buggert M, Gustafsson RK, Nielsen M, Lund O, Karlsson AC (2010) Interdisciplinary analysis of HIV-specific CD8+ T cell responses against variant epitopes reveals restricted TCR promiscuity. J Immunol 184(9):5383–5391

Hoof I, Peters B, Sidney J, Pedersen LE, Sette A, Lund O, Buus S, Nielsen M (2009) NetMHCpan, a method for MHC class I binding prediction beyond humans. Immunogenetics 61(1):1–13. doi:10.1007/s00251-008-0341-z

Jacob L, Vert JP (2008) Efficient peptide-MHC-I binding prediction for alleles with few known binders. Bioinformatics 24 (3):358–366

Jojic N, Reyes-Gomez M, Heckerman D, Kadie C, Schueler-Furman O (2006) Learning MHC I–peptide binding. Bioinformatics 22 (14):e227–e235

Lin HH, Ray S, Tongchusak S, Reinherz EL, Brusic V (2008) Evaluation of MHC class I peptide binding prediction servers: applications for vaccine research. BMC Immunol 9:8

Lundegaard C, Lamberth K, Harndahl M, Buus S, Lund O, Nielsen M (2008) *NetMHC*-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8–11. Nucleic Acids Res 36 (Web Server issue): W509-512.

Lundegaard C, Lund O, Buus S, Nielsen M (2010) Major histocompatibility complex class I binding predictions as a tool in epitope discovery. Immunology 130(3):309–318

Moutaftsi M, Peters B, Pasquetto V, Tscharke DC, Sidney J, Bui HH, Grey H, Sette A (2006) A consensus epitope prediction approach identifies the breadth of murine T(CD8+)-cell responses to vaccinia virus. Nat Biotechnol 24 (7):817–819

Nielsen M, Lundegaard C, Blicher T, Lamberth K, Harndahl M, Justesen S, Roder G, Peters B, Sette A, Lund O, Buus S (2007) NetMHCpan, a method for quantitative predictions of Peptide binding to any HLA-A and -B locus protein of known sequence. PLoS One 2(8).

Nielsen M, Lundegaard C, Worning P, Lauemoller SL, Lamberth K, Buus S, Brunak S, Lund O (2003) Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. Protein Sci 12(5):1007–1017

Peters B, Bui HH, Frankild S, Nielson M, Lundegaard C, Kostem E, Basch D, Lamberth K, Harndahl M, Fleri W, Wilson SS, Sidney J, Lund O, Buus S, Sette A (2006) A community resource benchmarking predictions of peptide binding to MHC-I molecules. PLoS Comput Biol 2(6):e65

Peters B, Sette A (2005) Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method. BMC Bioinforma 6:132

Rapin N, Hoof I, Lund O, Nielsen M (2010) The MHC motif viewer: a visualization tool for MHC binding motifs. Curr Protoc Immunol Chapter 18:Unit 18 17. doi:10.1002/0471142735.im1817s88

Robinson J, Waller MJ, Parham P, Bodmer JG, Marsh SG (2001) IMGT/HLA database—a sequence database for the human major histocompatibility complex. Nucleic Acids Res 29 (1):210–213

Stranzl T, Larsen MV, Lundegaard C, Nielsen M (2010) NetCTLpan: pan-specific MHC class I pathway epitope predictions. Immunogenetics 62(6):357–368. doi:10.1007/s00251-010-0441-4

Vita R, Zarebski L, Greenbaum JA, Emami H, Hoof I, Salimi N, Damle R, Sette A, Peters B (2010) The immune epitope database 2.0. Nucleic Acids Res 38(Database issue): D854–D862.

Wang P, Sidney J, Dow C, Mothe B, Sette A, Peters B (2008) A systematic assessment of MHC class II peptide binding predictions and evaluation of a consensus approach. PLoS Comput Biol 4(4):e1000048. doi:10.1371/journal.pcbi.1000048

Wang P, Sidney J, Kim Y, Sette A, Lund O, Nielsen M, Peters B (2010) Peptide binding predictions for HLA DR, DP and DQ molecules. BMC Bioinforma 11:568

Yu K, Petrovsky N, Schonbach C, Koh JY, Brusic V (2002) Methods for prediction of peptide binding to MHC molecules: a comparative study. Mol Med 8(3):137–148

Zhang H, Lund O, Nielsen M (2009a) The *PickPocket* method for predicting binding specificities for receptors based on receptor pocket similarities: application to MHC–peptide binding. Bioinformatics 25(10):1293–1299

Zhang H, Lundegaard C, Nielsen M (2009b) Pan-specific MHC class I predictors: a benchmark of HLA class I pan-specific prediction methods. Bioinformatics 25(1):83–89