Sequence analysis

Accurate approximation method for prediction of class I MHC affinities for peptides of length 8, 10 and 11 using prediction tools trained on 9mers

Claus Lundegaard*, Ole Lund and Morten Nielsen

Center for Biological Sequence Analysis – CBS, Department of Systems Biology, The Technical University of Denmark – DTU, Kemitorvet Build. 208, 2800 Lyngby, Denmark

Received on February 8, 2008; revised and accepted on April 4, 2008

Advance Access publication April 14, 2008

ABSTRACT

Summary: Several accurate prediction systems have been developed for prediction of class I major histocompatibility complex (MHC):peptide binding. Most of these are trained on binding affinity data of primarily 9mer peptides. Here, we show how prediction methods trained on 9mer data can be used for accurate binding affinity prediction of peptides of length 8, 10 and 11. The method gives the opportunity to predict peptides with a different length than nine for MHC alleles where no such peptides have been measured. As validation, the performance of this approach is compared to predictors trained on peptides of the peptide length in question. In this validation, the approximation method has an accuracy that is comparable to or better than methods trained on a peptide length identical to the predicted peptides.

Availability: The algorithm has been implemented in the web-accessible servers NetMHC-3.0: http://www.cbs.dtu.dk/services/NetMHC-3.0, and NetMHCpan-1.1: http://www.cbs.dtu.dk/services/NetMHCpan-1.1

Contact: lunde@cbs.dtu.dk

Supplementary information: Supplementary data are available at Bioinformatics online

1 INTRODUCTION

Determination of peptide binding to MHC class I is an important step in cytotoxic T cell lymphocyte (CTL) epitope discovery methods for class I MHC peptide binding. These methods have become increasingly accurate (Lundegaard et al., 2007; Moutaftsi et al., 2006; Peters et al., 2006), limiting the effort significantly. Most MHCs, however, prefer peptides of the length 9, making available binding data of 9mer peptides significantly more abundant than data for other lengths such as 8, 10 and 11mers as these more rarely binds to the MHCs. Since the amount of available data is crucial for the developing of accurate predictions (Yu et al., 2002), the number of accurate predictors of these other lengths is limited. Ligand motifs have been elucidated for several MHCs and groups of MHCs (Lund et al., 2004; Rammensee et al., 1999; Sette and Sidney, 1998). According to these motifs, the most important (anchor) residues are in general the positions 2, 3 and the C-terminal, disregarding the peptide length. However, for a limited number of non-human MHCs other peptide positions might be the primary anchors (see Supplementary Material). Using this knowledge, we exploited the possibility of generating pseudo 9mers from peptides of other length by fixating these positions and inserting or deleting residues at other positions. This resulted in a simple though remarkably accurate method to overcome the length problem using affinity predictions by 9mer predictors of such pseudo 9mers. This method will in principle work with any type of existing MHC 9mer binding prediction methods.

2 METHODS

2.1 9mer predictions

Here, we used predictions generated by NetMHC-3.0 (http://www.cbs.dtu.dk/services/NetMHC-3.0) (Buus et al., 2003; Nielsen et al., 2003). However, any 9mer:MHC binding prediction algorithm that accepts unknown amino acids (i.e. X) can be used.

2.2 Prediction of 8mer affinities

In 8mer peptides (e.g. EIGHTMER) an X is inserted repeatedly at either position 4, 5, 6, 7 or 8, resulting in five new pseudo peptides of length 9; EIGXHTMER, EIGHXTMER, EIGHTXMER, EIGHTMXER and EIGHTMEXR (Fig. 1A). The final predicted affinity is calculated as the geometrical mean of the five predicted affinities in nano Molar units.

2.3 Prediction of 10 and 11mer affinities

The longermers (e.g. TENELEVENS) are converted into 9mers by deleting 1 (10mers) or 2 (11mers) residues at positions 4, 5, 6, 7, 8 or 9, resulting in six new pseudo-peptides; TENLEVENS, TENELEVENS, TENELEENS, TENELEENS, TENELEENS and TENELEENS (Fig. 1B). The final predicted affinity is calculated as the geometrical mean of the six predicted affinities in nano Molar units (Fig. 1C).

2.4 Evaluation

The method was evaluated using peptide IC50 and Kd affinity data extracted from the web site of the Immune Epitope Database and Analysis resource (IEDB) (Sette et al., 2005). This resulted in 8mer, 9mer and 10mer evaluation data available at

*To whom correspondence should be addressed.
8mer data: 1975 measurements distributed on 35 MHC alleles. 10mer data: 13,507 measurements distributed on 31 MHC alleles. 11mer data: 181 measurements, distributed on 25 MHC alleles. We evaluated the accuracy by Pearson correlation coefficients (PCC), and area under receiver operating characteristic (ROC) curves (AUC) using a binding cutoff of 500 nM.

3 RESULTS

We predicted affinities for 8mer, 10mer and 11mer data using the approximation method described in Methods. The overall PCCs for all predictions within each dataset were 0.69, 0.73 and 0.74, respectively. The overall AUCs were 0.86, 0.87 and 0.89, respectively. To calculate an AUC value, we needed both negative (IC50 or Kd ≥ 500 nM) and positive (IC50 or Kd < 500 nM) affinity data, thus we removed alleles having only binders or non-binders from the evaluation. This led to 27 compared alleles on 8mer peptides. The resulting PCC values had a mean of 0.72. Acceptable AUC values (above 0.7) were obtained for 25 of the 27 covered alleles (Table 1). To evaluate the 10mer approximation, we calculated PCC and AUC values for 27 alleles. Using a 500 nM threshold, 26 of 27 alleles had AUCs above 0.7 (Table 1).

To compare the approximation method with specifically trained methods, we used artificial neural networks (ANNs) previously trained as described in (Nielsen et al., 2003) on 10mer data. For 10mers, 2037 new data points covering 16 alleles had become available since training of 10mer specific ANNs, available at http://www.cbs.dtu.dk/services/NetMHC-3.0/evalset_10mers.xls. AUC values were calculated for each allele using either ANNs trained on 10mers or the approximation method described here (Supplementary Fig. 1). For 12 of the 16 alleles the approximation method performed better than the 10mer trained ANNs (P < 0.01).

For the currently small number of alleles for which the primary peptide anchor position(s) are in positions 4–8 the approximation method will not work well. Examples of such alleles and how to identify these are described in the Supplementary Material.


The total number of 8mer and 10mer data, is given as #8mer data and #10mer data, respectively. The number of binding peptides (Kd < 500 nM) in the 8mer and 10mer datasets are given as #8mer binders and #10mer binders, respectively. NA: Value not calculated due to lack of data.

Conflict of Interest: none declared.

REFERENCES


