Published in final edited form as:
J Immunol Methods. 2011 November 30; 374(1-2): 26–34. doi:10.1016/j.jim.2010.10.011.

Prediction of epitopes using neural network based methods

Claus Lundegaard[*], **Ole Lund**, and **Morten Nielsen**
Center for Biological Sequence Analysis, DTU Systems Biology, Building 208, Technical University of Denmark, DK-2800 Lyngby, Denmark

Ole Lund: lund@cbs.dtu.dk; Morten Nielsen: mniel@cbs.dtu.dk

Abstract

In this paper, we describe the methodologies behind three different aspects of the *NetMHC* family for prediction of MHC class I binding, mainly to HLAs. We we have updated the prediction servers servers, *NetMHC-3.2*, *NetMHCpan-2.2*, and a new consensus method, *NetMHCcons*, which, in their previous versions, have been evaluated to be among the very best performing MHC:peptide binding predictors available. Here we describe the background for these methods, and the rationale behind the different optimisation steps implemented in the methods. We go through the practical use of the methods, which are publicly available in the form of relatively fast and simple web interfaces. Furthermore, we will review results optained in actual epitope discovery projects where previous implementations of the described methods have been used in the initial selection of potential epitopes. Selected potential epitopes were all evaluated experimentally using *ex vivo* assays.

Introduction

The triggering event in CD8+ T cell activation is the binding of the T Cell Receptor (TCR) to a Major Histocompatibility Complex (MHC) class I molecule, in complex with a peptide. However, in order to have the properties of an epitope, a given subpeptide must be processed from a larger polypeptide and must be able to bind to the gene product of a relevant MHC allele. This processing includes two major steps: proteasomal cleavage and binding to the Transporter associated with Antigen Presentation (TAP), (Stevanovic 2005). Not all theoretical subpeptides are created by these events, as both the constitutive proteasome and, to a larger extent, the immunoproteasome have protease activity with preferences for certain cleavage sites (Kesmir et al. 2002; Nielsen et al. 2005; Saxová et al. 2003). The processing is usually independent of a given individual's genotype as the genes expressing the molecules participating in the peptide processing are close to monomorphic in the human population. In contrast, MHC encoding genes are highly polymorphic and more than 2000 functional alleles of the Human Leucocyte Antigens (HLA), HLA-A and HLA-B, have now been identified according to the IMGT/HLA database Release 3.1.0, 16 July 2010 (http://www.ebi.ac.uk/imgt/hla/). A given MHC binds only to a very specific set of peptides; only 1 out of 200 random, naturally occurring peptides are able to bind (Yewdell and Bennink 1999). In addition, considering the limitations created by the antigen processing and the limited TCR repertoire, the final part of random peptides that end up

[*]Corresponding author: Claus Lundegaard, Ph.D, associate professor, The Technical University of Denmark - DTU, Department of systems biology, Center for biological sequence analysis - CBS, Kemitorvet, build. 208, DK-2800 KGS. LYNGBY, DENMARK, Tel: (+45) 45 25 24 84, Mobile:(+45) 21 90 07 67, Fax:(+45) 45 93 15 85, lunde@cbs.dtu.dk.

being immunogenic is approximately 1/1000 (Yewdell and Bennink 1999). A very large number of different alleles have been shown to cluster into supertypes according to the peptide binding capacity. This way MHC alleles that have a significant overlap in the peptide binding repertoire can be clustered into the same fuctional group or supertype (Hertz and Yanover 2007; Lund et al. 2004; Reche and Reinherz 2004; Sette and Sidney 1998; Sette and Sidney 1999). Without initial screening, all subpeptides of lengths 8 to 11 in a given polypeptide could be potential epitopes. This large number of potential epitopes, for even a single protein, has necessitated the development of experimental shortcuts. One common approach is to use larger overlapping peptides in order to scan for interesting antigens and often to identify responsive peptides. However, a significant number of peptides must still be produced and tested and the minimal/exact epitope is often not identified without additional experiments. These additional experiments are often performed with cells isolated from the blood of formerly or currently infected individuals and the biological material is usually available in only limited amounts. Furthermore, a large majority of the tested peptides in such blind scans test negative. To save time and resources, prediction systems have been developed to limit the number of experiments needed to identify epitopes in a given individual. These methods have been used in epitope discovery with significant success and now have a success rate of approximately 10%, as described previously (Lundegaard et al. 2010).

The most important event in the MHC class I epitope presentation pathway is the peptide binding to the MHC molecule (Yewdell and Bennink 1999) and considerable research has focused on predicting this specific event. Historically, the development of the most successful methods for MHC:peptide binding predictions has been closely connected to data generation. Such examples are SYFPEITHI which was developed on the basis of eluted peptides (Rammensee et al. 1995; Rammensee et al. 1997; Rammensee et al. 1999) and BIMAS which was developed using the stability measured as half life ($t_{1/2}$) (Parker et al. 1994; Parker et al. 1994). The type of the available data has naturally influenced the choice of methodology applied in the development of the different prediction methods. In the case of MHC:peptide binding, the event can be determined either directly by biochemical means or indirectly by cellular responses. For practical reasons, only the first type can be generated in amounts that enable the development of accurate prediction algorithms, but the latter has been used extensively for validation and as a supplement to biochemically identified data. Biochemically determined peptide binding data can, fundamentally, be obtained either by a direct measurement of the equilibrium constant or by identification of peptides bound by MHC. This creates two fundamentally different types of data, as the biochemically determined data has an exact value that can vary within the limits of the measurements, whereas the elution data is in a binary format (binder/non binder) and therefore only positive binding data can be directly detected. Non-binding peptides can only be indirectly deduced by their absence from a pool of eluted peptides where at least one peptide from the same host protein is present. As described in this paper, we use binary data obtained from the SYFPEITHI database (Rammensee et al. 1995; Rammensee et al. 1999) as well as peptides with a measured affinity for a given allele extracted from the Immune Epitope DataBase and analysis resource (IEDB) (Vita et al. 2010). In this paper, we additionally describe how to use MHC:peptide binding prediction servers and how to interpret the output.

MHC peptide binding predictions used in epitope discovery

As sequence data on genomic scale are growing rapidly, the usefulness of predictive systems is becomming more and more apparent. Screening for epitopes from the complete proteomes of smaller viruses such as HIV, HCV, or influenza A virus are possible using experimental epitope scanning techniques, e.g., overlapping peptides of length 15–20 (Kiepiela et al. 2007). But for more complex viruses such as smallpox or intracellular bacteria or parasites

as *Mycobacteria tuberculosis* or *Leishmania major*, respectively, a full proteome peptide scan is not experimentally feasible. However, it have been shown that MHC class I *in silico* models can significantly reduce the effort needed to conduct full genome epitope discovery experiments regarding pathogens with larger genomes such as smallpox and vaccinia (Moutaftsi et al. 2006). The importance of *in silico* methods in modern epitope discovery are increasingly emphasized and several large and important epitope discovery projects would have needed significantly more resources had they been carried out without the aid from these computational models (Lundegaard et al. 2010). *NetMHC* predictions have been used in a number of epitope discovery projects and here we will briefly summarize a few selected examples.

In order to identify relavant CD8+ T cell epitopes in Vaccinia virus the focus were put on epitopes that would be present not only in Vaccinia virus but also in additionally seven related pox strains (2 variola strains, 3 vaccina strains and 2 cowpox strains). The full genomes were scanned for conserved orthologue proteins (Tang et al. 2008) and the 157 identified proteins were then used as input to the *NetCTL-1.0* server (Larsen et al. 2005; Larsen et al. 2007). *NetCTL* integrates *NetMHC* peptide binding predictions with prediction of TAP transport (Peters et al. 2003) and predictions of appropriate proteasomal cleavage (Kesmir et al. 2002; Nielsen et al. 2005). Predicted epitopes that were 100% conserved in all seven strains were then considered, and 177 peptides were synthesized for use in experimental evaluation. The evaluation of the predicted epitopes was performed using peripheral blood mononuclear cells (PBMCs) extracted from buffy coats from healthy individuals. The blood donors were selected to be in an age group that would have been participating in the gereral pox vaccination program that ended in the 1970s. Thus the original immunizations were done more than 30 years before the test. Of the 177 peptides, eight (or 4.5%) were identified as CTL epitopes. The actual MHC binding was verified in biochemical assays and interestingly only peptides with a measured MHC affinity stronger than 5 nM were shown to be among the identified epitopes.

Another epitope discovery project concerned the human pathogen Influenza A, focusing on current H1N1 strains evolved from the 1918 Spanish flu. The genome derived proteomes of a large number of human H1N1 strains were scanned using the integrated CTL epitope prediction system *NetCTL* (Wang et al. 2007). The goal was to identify epitopes that were all conserved to a high extend and 15 potential epitopes should be selected restricted to each of the twelve considered supertypes (Lund et al. 2004). In order to maintain high conservation and at the same time select the desired number of peptides for eksperimental validation, the predicted binding affinity was not always as strong as the generally accepted threshold of 500 nM. Of the 180 predicted epitopes, 167 were synthesized for validation. The evaluation was performed using PBMCs from buffy coats from healthy blood donors in an age group with a high likeliness of having suffered from on average three Influenza A infections, thus having a high likeliness of having experienced a Influenza A H1N1 infection. Of the 167 peptides tested, 13 (8%) were shown to be able to induce a T cell response. These epitopes were all selected for being conserved in human H1N1 strains. However, for 11 of the 13 peptides 100 % identical matches existed in proteomes of two examined H1N5 bird flu strains (Wang et al. 2007), and all 13 epitopes were 100 % conserved in the proteome of the later 2009 H1N1 pandemic strain (unpublished results).

The two above examples describe searches for epitopes were vaccines or infections some time back had induced the original immune response. As an example on *NetMHC* directed epitope discovery considering ongoing infections, we will briefly summarize the result of an CD8+ T cell epitope discovery experiment concerning HIV infected individuals. Here, another approach for full genome variance coverage called *EpiSelect* (Perez et al. 2008) was used.

Potential epitopes was predicted from the in silico translated genomes isolated from more than 300 HIV strains of diverse subtypes using the integrated prediction system *NetCTL*. Each potential epitope were restricted to at least one of nine common HLA class I supertypes. For each of the nine supertypes peptides were iteratively selected by the *EpiSelect* scheme. According to this approach a new peptide restricted to a given supertype is preferred if it is present in the proteome of HIV strains not already targeted by previous selected peptides restricted to the same supertype. Of the selected peptides 184 were synthesized and tested against PBMCs from 31 HIV patients infected with various HIV subtypes. Of the tested 184 HLA class I supertype-restricted epitopes, as many as 114 (62%) were recognized by at least one study subject, and 45 of these were novel epitopes not previously reported in the literature.

As an example of successful genome wide epitope discovery projects using other MHC:peptide binding prediction methods than *NetMHC*, we have previously reviewed a large pox study by Moutaftsi et al. (Lundegaard et al. 2006; Moutaftsi et al. 2006) were T cell epitopes responsible for 95% of the total immunity were identified by MHC:peptide binding predictions. Here, we will shortly summarize a recent effort concerning the parasite *Leishmania major*, which proteome was mined with respect to Mouse MHC class I epitopes (Herrera-Najera et al. 2009). The authors describe that they did consensus epitope predictions of 8272 annotated protein sequences using several steps of predictions and filtering in order to limit the experiments needed for identification of verified epitopes. First prediction were performed restricted to the mouse alleles H-2D^{d} and H-2K^{d} of all possible octamer, nonamer, decamer, and hendecamer peptides from the full *Leishmenia major* proteome using the prediction system *RankPep*, that integrates MHC binding and predictions of proteasomal cleavage into a final prediction score (Reche et al. 2004). The top predictions of these were taken to next step where the peptides were predicted using five to eight different prediction tools, including SYFPEITHI (Rammensee et al. 1997) and BIMAS (Parker et al. 1994). Each of the peptides were now given a score based on the average predicted rank of the peptide compared with all other predicted peptides from the same protein. Finally the peptides were filtered for similarity to peptides from the Mouse host proteome, as well as the human proteome. This step was performed in order to avoid the potential induction of autoimmunity. 78 potential class I CD8 epitopes were identified. The 26 peptides that reached the best consensus rank score were tested for immunogenicity. The experimental validation was obtained by direct immunization with the peptides in a described vaccine formulation and 14 of the 26 (54%) turned out to be immunogenic in this setup. Here, was found that a relatively high proportion was immunogenic, which is in contrast to several other epitope discovery projects using *in silico* methods where approximately 10% of the tested peptides turned out to be immunogenic (Lundegaard et al. 2010). However, in the Herrera-Najera study was used direct immunization with peptides, were in most other projects it is tested if the peptide can recall a CTL response using monocytes from individuals having immune responses caused by an existing or previous infection with either the native or an attenuated form of the relevant pathogen.

Use of the web accessible prediction servers

NetMHC-3.2

The use and interpretation of the server output has been published for the previous version, *NetMHC-3.0*, which is close in functionality to the current server (Lundegaard et al. 2008). However, for completeness, we will briefly describe the use of the current server as well as the differences from *NetMHC-3.0*. *NetMHC-3.2* predicts the binding affinity of either a list of peptides with a defined length (8–11 residues) or all possible sub-peptides hosted within full-length proteins restricted to 57 human alleles and 22 animal alleles.

The input is taken as one or more protein sequences in FASTA format, each sequence not more than 20,000 amino acids in length and with a minimum length corresponding to the selected length of the predicted epitopes. Alternatively the input can be a raw list of peptides all with a uniform length equal to the selected length of the prdicted epitopes. One or more MHC alleles must selected. In *NetMHC-3.2* we have out-phased previous PSSM based predictors and all predictions obtained by the current server is now based on trained artificial neural networks (ANN), see the description of the algorithms later..

The output is displayed as raw text with a header indicating the server name and version, the first selected allele and the date followed by the prediction outputs in a column format. The order of the information in each row are the following: The position of the first amino acid of the peptide relative to the native polypeptide, the peptide sequence, the raw prediction score, the predicted affinity in nM units, and finally an indication if the peptide is predicted to be a strong binder (SB), i.e. binding with an affinity stronger than 50 nM (SB) or a weak binder (WB), binding stronger than 500 nM. As default the predictions are given in the order of appearance in the hosting polypeptide, however an optional sorting by predicted affinity can be requested before submission. In the output also exist a link for downloading the predictions in a tab-separated format, which can easily be opened by standard spreadsheet software.

New in *NetMHC-3.2* is that the prediction of decamer and hendecamer peptides are given as an average of the output from the approximation method and the output of direct predictions by ANN trained on exact length data if such network exists.

NetMHCpan-2.2

NetMHCpan-2.2 predicts the binding affinity of either a list of peptides with a defined length (8–11 residues) or all possible sub-peptides hosted within full-length proteins restricted to any known MHC molecule. The input of the proteins or peptides to be predicted as well as the selected length of peptides to be predicted is taken identically to *NetMHC-3.2* as described above. For selecting the restricting allele(s) one of three possibilities exists. Either use the scrolldown window after limiting the possible alleles, or type in a list of allele in the appropriate text field. As a final option an input can be taken a full MHC sequence in FASTA format if the given MHC allele does not exist in the selection lists. If a downloadable tab-formated output file is to be generated this must be indicated before submission of the prediction.

The output of *NetMHCpan-2.2* is similar to the output from *NetMHC-3.2* except that also a rank score (%Random) is given. The rank score is determined as the rank fractile of the given prediction score in a sorted list of prediction scores of 1,000,000 randomly selected naturally occurring peptides.

Description of the prediction algorithms

Training data

The data used for making supertype specific position specific scoring matrices (PSSMs) were eluted ligands and epitopes of length 9, 10 and 11 amino acids extracted from the SYFPEITHI database (Rammensee et al. 1999). We used the supertypes associations as defined by Lund et al. (Lund et al. 2004), and pooled all available peptides associated with an allele belonging to a given supertype and generated PSSMs as described below.

For training of the ANNs used in the MLI contest machine learning in immunology competition (MLI), URL:http://www.kios.org.cy/ICANN09/MLI.html, we used peptide data with an associated biochemically measured affinity extracted from the IEDB database

(www.immuneepitope.org) (Vita et al. 2010). We used data with affinity measures already publicly available at the time of training plus additional data on the way to be public. This dataset was kindly created and made available for us by Dr. Björn Peters. In total, we have used 102,146 peptide-affinity pairs covering 102 MHC alleles for this training.

Despite the fact that the raw number of data is large enough not to be a limitation regarding the alleles selected to be included in the MLI contest (HLA-A*0101, HLA-A*0201, and HLA-B*0702), it is a potential problem that the vast majority the affinity measurements in the databases has been made to peptides that were already suspected to bind at least one MHC molecule. Relatively few of the peptide:MHC binding data have been made by blind testing of all possible peptides covering a full antigen, which biases the available data against peptides that contain MHC binding patterns. However, it can be deduced from biological data that only a very limited part of all possible peptides will be able to bind to a given MHC with any significant strength. Thus any random peptide, can be considered a non-binder in relation to a particular MHC. This assumption has been successfully tested by adding a number of randomly selected, homology reduced naturally occurring peptides to the dataset and assign an affinity corresponding to the upper limit of the measurement range (50 μM), i.e., non binder. This approach was used in the training of both the *NetMHC-3.2* (n0001) and *NetMHCpan-2.2* (n0002) predictors used in the MLI competition.

Position specific scoring matrices in MHC ligand predictions

Due to the nature of the binding grove in MHC class I, peptides can be easily aligned within each length class. An obvious choice of method for generating position specific scoring matrices (PSSMs) from positive binding examples for a given MHC allele was the procedure including pseudo counts and sequence weighting developed for protein family identification and implemented in PsiBLAST (Altschul et al. 1997). The exact procedure is previously described (Nielsen et al. 2004) but is recaptured here for completeness:

1. Peptides of a uniform length known to bind to a specific MHC molecule is aligned by simple stacking
2. Sequences weighting using sequence clustering
3. Pseudo count correction
4. Weight on pseudo count

In our approach, we used ligand and epitope examples from the SYFPEITHI database. As the majority of the data is available as nonamers, we concentrated on making matrices for this length. However, as many alleles have a limited number of even nonamer peptides, we developed a scheme in order to benefit from the additional information in longer peptides. This is done by a so-called lmer approach were longer peptides are resized into nonamers and included in the statistics. Since it has been shown that for most HLA class I alleles, the amino acid residues positioned in the peptide position 2, to some degree position 3, and the C-terminal position are most important for determining the binding to the HLA molecule, these positions were always kept and the resized peptides were constructed by in all peptides subsequently removing one (length 10) or two (length 11) consecutive residues at positions P4 to P(L-1) where L is the length of the peptide (see figure 1).

Allele specific predictions by the NetMHC-3.2 method (n00001)

Several experimental high-through-put methods are now able to generate MHC:peptide affinity data in large amounts (Harndahl et al. 2009; Sidney et al. 2001; Sylvester-Hvid et al. 2002) and we have been developing systems to be able to accurately predict these values. Since we have a long running experience in employing ANNs for biological prediction systems this was our first choice of method for this task (Nielsen et al. 2003). Several other

scientific groups have successfully used this method for epitope predictions (Adams and Koziol 1995; Bhasin and Raghava 2004; Buus et al. 2003; Ramakrishna et al. 1997) even though not all ANN based methods are superior to the more refined linear methods (Peters et al. 2006). A number of different MHC class I peptide binding prediction methods have been recently reviewed (Lafuente and Reche 2009; Lundegaard et al. 2010; Lundegaard et al. 2010; Toussaint and Kohlbacher 2009; Yang and Yu 2009)

We use a standard feed forward network with back propagation as previously described (Lund et al. 2005; Nielsen et al. 2003). For peptide input, we use two different approaches (Figure 2). Either sparse encoding, where each amino acid is represented by a vector of twenty nodes, one having the value 0.9 and the 19 other having the value 0.05. The position with the value 0.9 is unique for each of the twenty standard amino acids. The second approach is defined using the BLOSUM50 matrix (Henikoff and Henikoff 1992) as taken from the NCBI repository. Here, the twenty BLOSUM substitution scores represent each amino acid. For a nonamer peptide, this will with both approaches result in an input vector of 180 nodes. To include the information from the ligand data, which have no measured affinity, we created PSSMs with the lmer approach described above and used the 9 position scores for a given peptide as additional input to the ANN, resulting in a total of 189 input nodes for a nonamer peptide. For this purpose we use supertype specific matrices. The ANNs have one single output node. We trained on the measured K_d values transformed using equation 1 in order to have output values in the range 0.0–1.0. We trained the ANN to minimize the error (sum of squared errors) between the output and the log-transformed affinity using a 5-fold cross validation approach using 4/5 of the peptides to optimize the weights and 1/5 as test set to stop the network training and avoid overtraining. For each cross-validation partition, we train two parallel modes with either sparse or BLOSUM50 encoded peptide input. For both modes we are using 1, 2, 4, 32 and 64 nodes in the hidden layer. After training, the architecture giving the best test performance is saved giving two ANNs (1 sparse and 1 BLOSUM50) for each partition. As this is not an appropriate way to obtain a cross validated performance because of potential overfitting we generally use external validation sets kept out of the train/test cycles to estimate the predictive performance. The final predictor is an ensemble of 10 (5 partitions X 2 encoding schemes) ANNs. The final output from a given predictor is calculated as the simple mean of the 10 outputs.

$$S = 1 - log_{50000}(K_d) \qquad \text{Equation 1}$$

where S is the output score and K_d is given in nM units.

The number of decamer peptide affinity data is generally much smaller than for nonamer peptides, which is a problem for generation of accurate prediction systems. Moderately accurate PSSMs can be created using the information from just a very few data points ([NO STYLE for: Lundegaard 2004]). However, for the more accurate ANN systems here described we need at least 100 data points to learn the peptide binding properties of a given HLA molecule (Yu et al. 2002). This fact motivated us to test if it is possible to utilize the ANNs trained on nonamer data to predict peptides of length 10 and 11. This approach has previously been published (Lundegaard et al. 2008), but is described here for completeness. As described earlier, the positions 1–3 and the N-terminal position are the most important for binding. We therefore made pseudo nonamers from the decamer peptides using an approach similar to the lmer approach for using longer peptides in the generation of PSSMs of length 9 described under PSSMs. In order to use nonamer trained predictors, we first convert the longer peptide into a nonamer peptide. For decamers this can be done as

previously described by removing in turn the amino acid at positions 4 to 8 creating 6 new pseudo peptides of length nine. The affinity of each of the six peudo-nonamers was next predicted using the convential 9mer ANN. The final approximation affinity prediction for the decamer was finally calculated as the geometrical mean of the six predicted affinities. This has turned out to be a good approach, especially in cases were the number of actual decamer data is very small. For some alleles, we did have enough data to train specific decamer ANNs, but using previously published iTopia measurements as an independent test set (Lin et al. 2008), we found that a simple mean of the outputs from the approximation approach and decamer trained predictors was even more accurate than any of the two approaches separately. Thus this average is the prediction used for decamer predictions in the MLI contest for the NetMHC method, and is also used in *NetMHC-3.2* web accessible server where decamer trained ANNs are available.

NetMHC-3.0 and the method behind have been independently benchmarked on different evaluation data and have in these benchmarks always had a predictive performance superior to the compared methods on the employed datasets (Lin et al. 2008; Peters et al. 2006).

The IEDB implementation of the *NetMHC* predictor is called *ANN* in the IEDB framework (Vita et al. 2010). However, *ANN* does not use the approximation method and thus does not offer decamer predictions for the alleles were decamer predictors could not be generated due to lack of training data.

Pan-specific predictions (NetMHCpan-1.2/n0002)

The large majority of HLA-A and -B alleles have never been investigated in relation to peptide binding. This is a major challenge regarding the goal of being able to predict peptide binding to any HLA-A or -B allele. To go beyond the allele-specific approach without requiring peptide data specific for each allele in question, we have earlier developed a so-called pan-specific MHC binding method that allows for prediction of peptide binding to any MHC molecules of known protein sequence (Hoof et al. 2009; Nielsen et al. 2008). Here we give a short description of rational behind and implementation of this method. When looking at crystal structures of peptide:MHC complexes, it becomes apparent that most of the polymorphic residues of HLA alleles are placed at positions in contact with binding peptides (Nielsen et al. 2008) (Figure 3). This is also expected to be the positions having the most effect on peptide binding. Using structural data, we identified HLA residues in contact with bound peptides and subsequently we checked which of these that were polymorphic in known functional HLA-A, -B, or -C alleles. This revealed 34 polymorphic residues that we then presented as so-called pseudo-sequences (Figure 3). It was then possible to train ANNs using these pseudo-sequences as input paired with a given nonamer peptide with known affinity. Thus the neural network was trained to output the affinity of a given MHC:peptide pair having the MHC represented by the pseudo sequence. Like in the allele specific training (*NetMHC-3.2*) we used a five fold cross validation scheme using both BLOSUM50 and sparse encoding (see the previous section), but no PSSM input was applied in this construction. This architecture results in an input layer of 880 input nodes for a nonamer peptide pseudo-sequence pair. The same type of feed forward ANN with back-propagation were used as in the case of allele specific ANNs.

These types of pan-specific ANNs have turned out to be very successful in predicting peptide affinity both to alleles with no available training examples, but also for alleles characterized by very few data points (Hoof et al. 2009; Nielsen et al. 2007). In a benchmark study, the *NetMHCpan-1.0* ANN based predictor were compared with other pan predictors using a large independent evaluation set (Zhang et al. 2009). Often data points are selected for measurement based on prediction systems and these same systems may be doing artificially well on such data as at least all positive data were already predicted positive by

the systems. To avoid this bias the evaluation benchmark set were depleted from data that had been selected for experimental validation due to positive predictions by any *NetMHC* predictor. The results from this evaluation showed that the *NetMHCpan-1.0* predictor is outperforming all other pan-specific predictors in this benchmark. The performance of *NetMHCpan-1.0* was also compared to the allele specific trained *NetMHC-3.0*. *NetMHCpan-1.0* performed nearly as good as *NetMHC-3.0* when evaluated on alleles, which had large peptide coverage in the training sets, and performed better than *NetMHC-3.0* for allele characterized by few training examples.

A consensus method (NetMHcCons/n00003)

In the above described benchmark study (Zhang et al. 2009), the pan-approach was nearly as good as the allele specific training procedure and even though the latter was the best performing for most alleles, the pan-specific method turned out to be the best performing for several other alleles covered by limited peptide binding data for training. As consensus methods have also in MHC:peptide binding being considered to be superior to single method predictions (Flower 2003; Moutaftsi et al. 2006; Trost et al. 2007), the simple mean of the predicted affinity was used as a consensus prediction in the benchmark. This simple mean of the two predictions turned out to outperform both *NetMHCpan-1.0* and *NetMHC-3.0* thus a consensus method (*NetMHCCons*) giving as output a simple mean of the prediction scores of *NetMHCpan-2.2* and *NetMHC-3.2* was participating as n00003 in the MLI competition.

Discussion

In a number of benchmarks *NetMHC* and *NetMHCpan* have been shown to have a competing edge to other prediction systems. This goes both when testing several algorithmic methods using identical training sets (Peters et al. 2006), as well as testing the finally trained methods on external evaluation sets, as is the case in the present competition (Lin et al. 2008; Roomp et al. 2010; Zhang et al. 2009). The latest comparison results that confirmed superior prediction accuracy are reported elsewhere in this issue (Zhang et al, 2011). The conclusion, also from the latest true blind assessment, is that no large progress has been gained regarding the methods of MHC class I peptide predictions for the last years. Nonetheless efforts like the MLI competition are essential as they might inspire for new approaches. Also, as this competition shows, the best servers can today predict the peptide binding quite accurately for the included alleles which are all well studied, and we can only hope for a new competition regarding not only less well studied HLA alleles but also including MHCs from some of the important animal models and agriculturally important animal species in order to examine the potential of the more general methods. Such larger dataset would further be of interest, especially containing datapoints not selected by any prediction. A full scan of one or two smaller proteins would be an ideal dataset

We have also presented some examples of the successful use of in silico methods for large scale epitope discovery showing some of the great potential researchers have using these methods to reducing experiment costs and efforts. This competition puts focus on the binding prediction accuracy but it is of course the usefulness in experimental science that is the real test for the acid of predictive systems in epitope discovery. Some ill-performed experiments concluding that binding predictions did not work (Andersen et al. 2000) were keeping experimentalists from benefiting of the developed methods. Today, however, most experimental scientists have been convinced by the amount of documentation. We now face another kind of suspicion by some experimentalists claiming that MHC restriction in many cases fails to explain CTL responses against specific peptides (Altfeld et al. 2005). This, however, in the far majority of cases results from a misunderstanding of the supertype concept, that all alleles assigned to a given supertype will bind any peptide restricted to another allele from the same supertype. In fact, it can been shown that the large majority of

the responses, that seem to lack HLA restriction, can be explained when using allele specific predictions for the fully-typed HLA-types of the donor (Hoof et al. 2010). As it is now possible to predict binding affinities for specific alleles defining a population, we might benefit from using such predictions in order to select pools of peptides with broad population coverage as an alternative to the supertype approach.

Acknowledgments

We thank Colleen Ussery for suggestions regarding improvement of the written language.

References

Adams HP, Koziol JA. Prediction of binding to MHC class I molecules. J Immunol Methods. 1995; 185:181–190. [PubMed: 7561128]

Altfeld M, Allen TM, Kalife ET, Frahm N, Addo MM, Mothe BR, Rathod A, Reyor LL, Harlow J, Yu XG, Perkins B, Robinson LK, Sidney J, Alter G, Lichterfeld M, Sette A, Rosenberg ES, Goulder PJ, Brander C, Walker BD. The majority of currently circulating human immunodeficiency virus type 1 clade B viruses fail to prime cytotoxic T-lymphocyte responses against an otherwise immunodominant HLA-A2-restricted epitope: implications for vaccine design. J Virol. 2005; 79:5000–5005. [PubMed: 15795285]

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25:3389–3402. [PubMed: 9254694]

Andersen MH, Tan L, Søndergaard I, Zeuthen J, Elliott T, Haurum JS. Poor correspondence between predicted and experimental binding of peptides to class I MHC molecules. Tissue Antigens. 2000; 55:519–531. [PubMed: 10902608]

Bhasin M, Raghava GP. Prediction of CTL epitopes using QM, SVM and ANN techniques. Vaccine. 2004; 22:3195–3204. [PubMed: 15297074]

Buus S, Lauemoller SL, Worning P, Kesmir C, Frimurer T, Corbet S, Fomsgaard A, Hilden J, Holm A, Brunak S. Sensitive quantitative predictions of peptide-MHC binding by a 'Query by Committee' artificial neural network approach. Tissue Antigens. 2003; 62:378–384. [PubMed: 14617044]

Flower DR. Towards in silico prediction of immunogenic epitopes. Trends Immunol. 2003; 24:667–674. [PubMed: 14644141]

Harndahl M, Justesen S, Lamberth K, Røder G, Nielsen M, Buus S. Peptide binding to HLA class I molecules: homogenous, high-throughput screening, and affinity assays. J Biomol Screen. 2009; 14:173–180. [PubMed: 19196700]

Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA. 1992; 89:10915–10919. [PubMed: 1438297]

Herrera-Najera C, Piña-Aguilar R, Xacur-Garcia F, Ramirez-Sierra MJ, Dumonteil E. Mining the Leishmania genome for novel antigens and vaccine candidates. Proteomics. 2009; 9:1293–1301. [PubMed: 19206109]

Hertz T, Yanover C. Identifying HLA supertypes by learning distance functions. Bioinformatics. 2007; 23:e148–e155. [PubMed: 17237084]

Hoof I, Peters B, Sidney J, Pedersen LE, Sette A, Lund O, Buus S, Nielsen M. NetMHCpan, a method for MHC class I binding prediction beyond humans. Immunogenetics. 2009; 61:1–13. [PubMed: 19002680]

Hoof I, Pérez CL, Buggert M, Gustafsson RK, Nielsen M, Lund O, Karlsson AC. Interdisciplinary Analysis of HIV-Specific CD8+ T Cell Responses against Variant Epitopes Reveals Restricted TCR Promiscuity. J Immunol. 2010

Kesmir C, Nussbaum AK, Schild H, Detours V, Brunak S. Prediction of proteasome cleavage motifs by neural networks. Protein Eng. 2002; 15:287–296. [PubMed: 11983929]

Kiepiela P, Ngumbela K, Thobakgale C, Ramduth D, Honeyborne I, Moodley E, Reddy S, de Pierres C, Mncube Z, Mkhwanazi N, Bishop K, van der Stok M, Nair K, Khan N, Crawford H, Payne R, Leslie A, Prado J, Prendergast A, Frater J, McCarthy N, Brander C, Learn GH, Nickle D,

Rousseau C, Coovadia H, Mullins JI, Heckerman D, Walker BD, Goulder P. CD8+ T-cell responses to different HIV proteins have discordant associations with viral load. Nat Med. 2007; 13:46–53. [PubMed: 17173051]

Lafuente EM, Reche PA. Prediction of MHC-peptide binding: a systematic and comprehensive overview. Curr Pharm Des. 2009; 15:3209–3220. [PubMed: 19860671]

Larsen MV, Lundegaard C, Lamberth K, Buus S, Brunak S, Lund O, Nielsen M. An integrative approach to CTL epitope prediction: a combined algorithm integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions. Eur J Immunol. 2005; 35:2295–2303. [PubMed: 15997466]

Larsen MV, Lundegaard C, Lamberth K, Buus S, Lund O, Nielsen M. Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction. BMC Bioinformatics. 2007; 8:424. [PubMed: 17973982]

Lin HH, Ray S, Tongchusak S, Reinherz EL, Brusic V. Evaluation of MHC class I peptide binding prediction servers: applications for vaccine research. BMC Immunol. 2008; 9:8. [PubMed: 18366636]

Lund, O.; Nielsen, M., et al. Immunological Bioinformatics. The MIT Press; Cambridge, Massachusetts, London, England: 2005.

Lund O, Nielsen M, Kesmir C, Petersen AG, Lundegaard C, Worning P, Sylvester-Hvid C, Lamberth K, Røder G, Justesen S, Buus S, Brunak S. Definition of supertypes for HLA molecules using clustering of specificity matrices. Immunogenetics. 2004; 55:797–810. [PubMed: 14963618]

Lundegaard C, Hoof I, Lund O, Nielsen M. State of the art and challenges in sequence based T-cell epitope prediction. Immunome Res. 2010; 6:S3. [PubMed: 21067545]

Lundegaard C, Lamberth K, Harndahl M, Buus S, Lund O, Nielsen M. 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. 2008; 36:W509–W512. [PubMed: 18463140]

Lundegaard C, Lund O, Buus S, Nielsen M. Major histocompatibility complex class I binding predictions as a tool in epitope discovery. Immunology. 2010; 130:309–318. [PubMed: 20518827]

Lundegaard C, Lund O, Nielsen M. Accurate approximation method for prediction of class I MHC affinities for peptides of length 8, 10 and 11 using prediction tools trained on 9mers. Bioinformatics. 2008; 24:1397–1398. [PubMed: 18413329]

Lundegaard C, Nielsen M, Lund O. The validity of predicted T-cell epitopes. Trends Biotechnol. 2006; 24:537–538. [NO STYLE for: Lundegaard 2004]. [PubMed: 17045685]

Moutaftsi M, Peters B, Pasquetto V, Tscharke DC, Sidney J, Bui HH, Grey H, Sette A. A consensus epitope prediction approach identifies the breadth of murine T(CD8+)-cell responses to vaccinia virus. Nat Biotechnol. 2006; 24:817–819. [PubMed: 16767078]

Nielsen M, Lundegaard C, Blicher T, Lamberth K, Harndahl M, Justesen S, Røder G, Peters B, Sette A, Lund O, Buus S. NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. PLoS ONE. 2007; 2:e796. [PubMed: 17726526]

Nielsen M, Lundegaard C, Blicher T, Peters B, Sette A, Justesen S, Buus S, Lund O. Quantitative predictions of peptide binding to any HLA-DR molecule of known sequence: NetMHCIIpan. PLoS Comput Biol. 2008; 4:e1000107. [PubMed: 18604266]

Nielsen M, Lundegaard C, Lund O, 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. [PubMed: 15744535]

Nielsen M, Lundegaard C, Worning P, Hvid CS, Lamberth K, Buus S, Brunak S, Lund O. Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach. Bioinformatics. 2004; 20:1388–1397. [PubMed: 14962912]

Nielsen M, Lundegaard C, Worning P, Lauemøller SL, Lamberth K, Buus S, Brunak S, Lund O. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. Protein Sci. 2003; 12:1007–1017. [PubMed: 12717023]

Parker KC, Bednarek MA, Coligan JE. Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. J Immunol. 1994; 152:163–175. [PubMed: 8254189]

Parker KC, Biddison WE, Coligan JE. Pocket Mutations of HLA-B27 Show That Anchor Residues Act Cumulatively To Stabilize Peptide Binding. Biochemistry. 1994; 33:7736–7743. [PubMed: 8011638]

Perez CL, Larsen MV, Gustafsson R, Norstrom MM, Atlas A, Nixon DF, Nielsen M, Lund O, Karlsson AC. Broadly immunogenic HLA class I supertype-restricted elite CTL epitopes recognized in a diverse population infected with different HIV-1 subtypes. J Immunol. 2008; 180:5092–5100. [PubMed: 18354235]

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. A community resource benchmarking predictions of peptide binding to MHC-I molecules. PLoS Comput Biol. 2006; 2:e65. [PubMed: 16789818]

Peters B, Bulik S, Tampe R, Van Endert PM, Holzhütter HG. Identifying MHC class I epitopes by predicting the TAP transport efficiency of epitope precursors. J Immunol. 2003; 171:1741–1749. [PubMed: 12902473]

Ramakrishna V, Negri DR, Brusic V, Fontanelli R, Canevari S, Bolis G, Castelli C, Parmiani G. Generation and phenotypic characterization of new human ovarian cancer cell lines with the identification of antigens potentially recognizable by HLA-restricted cytotoxic T cells. Int J Cancer. 1997; 73:143–150. [PubMed: 9334822]

Rammensee, HG.; Bachmann, J.; Stevanovic, S. MHC ligands and Peptide Motifs. Chapman & Hall; New York: 1997.

Rammensee HG, Friede T, Stevanoviíc S. MHC ligands and peptide motifs: first listing. Immunogenetics. 1995; 41:178–228. [PubMed: 7890324]

Rammensee H, Bachmann J, Emmerich NP, Bachor OA, Stevanovic S. SYFPEITHI: database for MHC ligands and peptide motifs. Immunogenetics. 1999; 50:213–219. [PubMed: 10602881]

Reche, PA.; Reinherz, EL. Artificial Immune Systems, Proceedings. Springer Verlag; Berlin: 2004. Definition of MHC supertypes through clustering of MHC peptide binding repertoires; p. 189-196.

Reche PA, Glutting JP, Zhang H, Reinherz EL. Enhancement to the RANKPEP resource for the prediction of peptide binding to MHC molecules using profiles. Immunogenetics. 2004; 56:405–419. [PubMed: 15349703]

Roomp K, Antes I, Lengauer T. Predicting MHC class I epitopes in large datasets. BMC Bioinformatics. 2010; 11:90. [PubMed: 20163709]

Saxová P, Buus S, Brunak S, Kesmir C. Predicting proteasomal cleavage sites: a comparison of available methods. Int Immunol. 2003; 15:781–787. [PubMed: 12807816]

Sette A, Sidney J. HLA supertypes and supermotifs: a functional perspective on HLA polymorphism. Curr Opin Immunol. 1998; 10:478–482. [PubMed: 9722926]

Sette A, Sidney J. Nine major HLA class I supertypes account for the vast preponderance of HLA-A and -B polymorphism. Immunogenetics. 1999; 50:201–212. [PubMed: 10602880]

Sidney J, Southwood S, Oseroff C, del Guercio MF, Sette A, Grey HM. Measurement of MHC/peptide interactions by gel filtration. Curr Protoc Immunol. 2001; Chapter 18(Unit 18.3)

Stevanovic S. Antigen processing is predictable: From genes to T cell epitopes. Transpl Immunol. 2005; 14:171–174. [PubMed: 15982559]

Sylvester-Hvid C, Kristensen N, Blicher T, Ferre H, Lauemoller SL, Wolf XA, Lamberth K, Nissen MH, Pedersen LO, Buus S. Establishment of a quantitative ELISA capable of determining peptide-MHC class I interaction. Tissue Antigens. 2002; 59:251–258. [PubMed: 12135423]

Tang ST, Wang M, Lamberth K, Harndahl M, Dziegiel MH, Claesson MH, Buus S, Lund O. MHC-I-restricted epitopes conserved among variola and other related orthopoxviruses are recognized by T cells 30 years after vaccination. Arch Virol. 2008; 153:1833–1844. [PubMed: 18797815]

Toussaint NC, Kohlbacher O. Towards in silico design of epitope-based vaccines. Expert Opin Drug Discov. 2009:4697.

Trost B, Bickis M, Kusalik A. Strength in numbers: achieving greater accuracy in MHC-I binding prediction by combining the results from multiple prediction tools. Immunome Res. 2007; 3:5. [PubMed: 17381846]

Vita R, Zarebski L, Greenbaum JA, Emami H, Hoof I, Salimi N, Damle R, Sette A, Peters B. The immune epitope database 2.0. Nucleic Acids Res. 2010; 38:D854–D862. [PubMed: 19906713]

Wang M, Lamberth K, Harndahl M, Røder G, Stryhn A, Larsen MV, Nielsen M, Lundegaard C, Tang ST, Dziegiel MH, Rosenkvist J, Pedersen AE, Buus S, Claesson MH, Lund O. CTL epitopes for influenza A including the H5N1 bird flu; genome-, pathogen-, and HLA-wide screening. Vaccine. 2007; 25:2823–2831. [PubMed: 17254671]

Yang X, Yu X. An introduction to epitope prediction methods and software. Rev Med Virol. 2009; 19:77–96. [PubMed: 19101924]

Yewdell JW, Bennink JR. Immunodominance in major histocompatibility complex class I-restricted T lymphocyte responses. Annu Rev Immunol. 1999; 17:51–88. [PubMed: 10358753]

Yu K, Petrovsky N, Schonbach C, Koh JY, Brusic V. Methods for prediction of peptide binding to MHC molecules: a comparative study. Mol Med. 2002; 8:137–148. [PubMed: 12142545]

Zhang H, Lundegaard C, Nielsen M. Pan-specific MHC class I predictors: a benchmark of HLA class I pan-specific prediction methods. Bioinformatics. 2009; 25:83–89. [PubMed: 18996943]

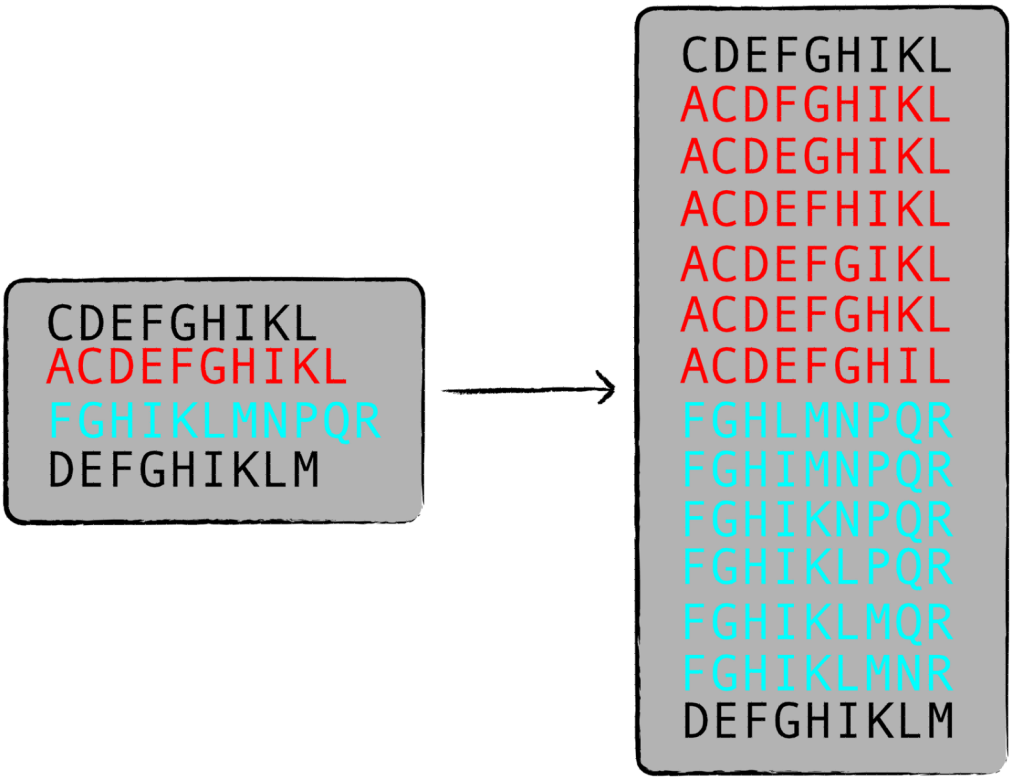

Figure 1.
Conversion of longer peptides to nonamers. Peptides longer than 9 get removed one more consecutive amino acids to a final length of nine. Positions P4 to P(L-1) are removed, where L are the length of the peptides, resulting in six new nonamer peptides for each longer peptide. The 6 new nonamer peptides are same color as the parent longer peptide.

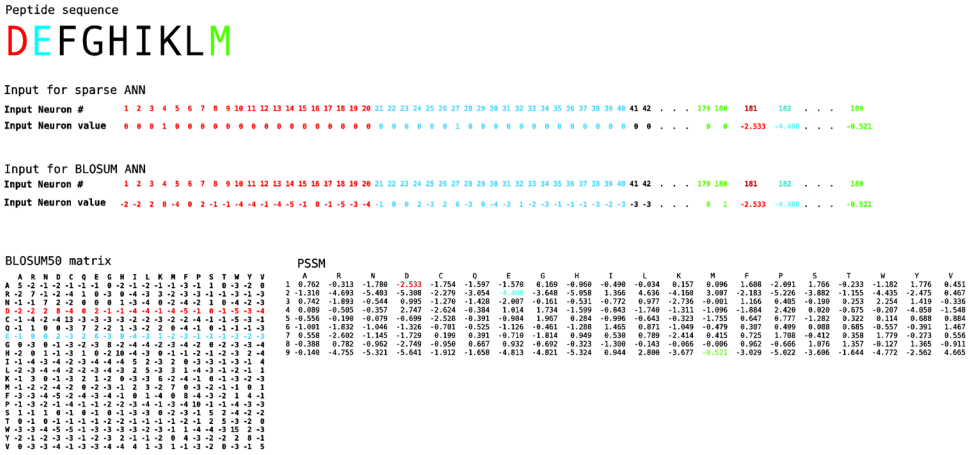

Figure 2.

Input encoding. For nonamer peptide inputs the input layer consist of 189 input units where the 180 first correspond to the peptide in sparse- or BLOSUM50 encoding, respectively. The last 9 units are taken from the PSSM generated using SYFPEITHI peptides. The input units are colorcoded according to the corresponding position in the input peptide.

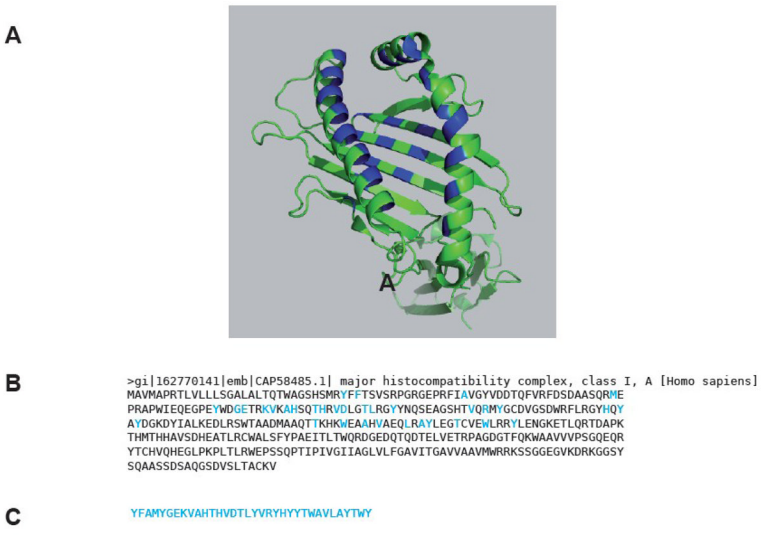

Figure 3.

Panel A shows a cartoon of the chrystal structure of HLA-A*0201 PDB entry 3HPJ. Polymorphic positions in contact distance from the peptide are color coded blue. Panel B shows the amino acid sequence of HLA-A*0201. Polymorphic positions in contact distance from the peptide are color coded blue.

Panel C shows the extracted pseudo sequence.