BIOINFORMATICS REVIEW Vol. 23 no. 24 2007, pages 3265–3275
doi:10.1093/bioinformatics/btm471

*Sequence analysis*

# Modeling the adaptive immune system: predictions and simulations

Claus Lundegaard[1,*], Ole Lund[1], Can Keşmir[1,2], Søren Brunak[1] and Morten Nielsen[1]

[1]Center for biological sequence analysis, CBS, Kemitorvet 208, Technical University of Denmark, DK-2800 Lyngby, Denmark and [2]Theoretical biology/bioinformatics, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands

Received on June 21, 2007; revised and accepted on September 10, 2007

Associate Editor: Jonathan Wren

**ABSTRACT**

**Motivation:** Immunological bioinformatics methods are applicable to a broad range of scientific areas. The specifics of how and where they might be implemented have recently been reviewed in the literature. However, the background and concerns for selecting between the different available methods have so far not been adequately covered..

**Summary:** Before using predictions systems, it is necessary to not only understand how the methods are constructed but also their strength and limitations. The prediction systems in humoral epitope discovery are still in their infancy, but have reached a reasonable level of predictive strength. In cellular immunology, MHC class I binding predictions are now very strong and cover most of the known HLA specificities. These systems work well for epitope discovery, and predictions of the MHC class I pathway have been further improved by integration with state-of-the-art prediction tools for proteasomal cleavage and TAP binding. By comparison, class II MHC binding predictions have not developed to a comparable accuracy level, but new tools have emerged that deliver significantly improved predictions not only in terms of accuracy, but also in MHC specificity coverage. Simulation systems and mathematical modeling are also now beginning to reach a level where these methods will be able to answer more complex immunological questions.

**Contact:** lunde@cbs.dtu.dk

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 INTRODUCTION

### 1.1 Immunology

The adaptive immune system of vertebrates is thought to be only 400 million years old and exists in most fish, amphibians, reptiles, birds and mammals (Thompson, 1995). Adaptive immunity is induced by lymphocytes and can be classified into two types: humoral immunity, mediated by antibodies, which are secreted by B lymphocytes and can neutralize pathogens outside the cells; and cellular immunity, mediated by T lymphocytes that eliminate infected or malfunctioning cells, and provide help to other immune responses. Diversity is the hallmark of the adaptive immune systems. Both the B and T lymphocyte-specific receptors for antigen recognition are assembled from variable (V), diversity (D), and joining (J) gene segments early in the lymphocyte development. There are multiple copies of V, D and J segments, and a huge repertoire of T and B cells is generated by the recombination of these segments, reviewed by Li *et al*. (2004). Another task faced by the immune system is the tolerance to self, which is handled by continuously removing receptors that react to self-epitopes.

Special immunoglobulin molecules (antibodies) mediate the humoral response. As mentioned above, the antibodies are produced by B lymphocytes that bind to antigens by their immunoglobulin receptors, which is a membrane bound form of the antibodies. When the B lymphocytes become activated, they start to secrete the soluble form of this receptor in large amounts. The antibody is Y-shaped, and each of the two branches functions independently and can be recombinantly produced and is then known as Fabs. The highly variable tip of the Fab, which can bind to epitopes is called the paratope and is made up of the so-called complementary determining regions (CDRs). Antibodies can coat the surface of an antigen such as a virus, so that it cannot function or infect cells, reviewed by Burton (2002). Antibody-covered viruses or bacteria are easily phagocytosed and destroyed by scavenger cells of the immune system, e.g. the macrophages. Antigenic proteins can be recognized by the antibodies in their native form without any cleavage or interactions with other molecules. Thus the humoral immune response reacts to extracellular pathogens, and the response is crucial in the defense against most pathogens.

B-cell epitopes are normally classified into two groups: continuous and discontinuous epitopes. A continuous epitope, (also called a sequential or linear epitope) is a short peptide fragment in a protein that is recognized by antibodies specific for that protein. A discontinuous epitope is composed of residues that are not adjacent in the primary structure (amino acid sequence), but are brought into proximity by the folding of the polypeptide. The classification is not clear-cut as discontinuous epitopes may contain linear stretches of amino acids, and continuous epitopes may show conformational preferences.

The cellular arm of the immune system consists of two parts; cytotoxic T lymphocytes (CTL), and helper T lymphocytes (HTLs). CTLs destroy cells that present non-self peptides (epitopes). HTLs are needed for B cells activation

*To whom correspondence should be addressed.

and proliferation to produce antibodies against a given antigen. CTLs on the other hand perform surveillance of the host cells, and recognize and kill infected cells, generally explained in Janeway *et al*. (2001). Both CTL and HTL are raised against peptides that are presented to the immune cells by major histocompatibility complex (MHC) molecules, which are the most polymorphic of mammalian proteins. The human versions of MHCs are referred to as the human leucocyte antigen (HLA). The cells of an individual are constantly screened for such peptides by the cellular arm of the immune system. In the MHC class I pathway, class I MHCs presents endogenous antigens to T cells carrying the CD8 receptor (CD8+ T cells). To be presented, a precursor peptide is normally first generated by the large cytosomal protease complex called the proteasome (Loureiroa and Ploegha, 2006). Generally, it then binds to the transporter associated with antigen processing (TAP) for translocation into the endoplasmic reticulum (ER), reviewed by Abele and Tampé (2004), but some peptides can enter the ER independently of TAP. This should be considered when dealing with virus-infected cells or tumors cells that might have reduced or absent TAP function. There are several ways that the peptide can enter the ER without TAP function depending on the origin and properties of the peptide. The most well-established model, however, is for proteins containing a signal peptide. Such proteins are translated directly into the ER through the Sec61 transporter complex and sometimes the cleaved-off signal peptide will end up in ER. This model is especially relevant for peptides binding to HLAs belonging to the abundant A2 HLA serotype where TAP-independent presentation is responsible for up to 10% of the A2 restricted epitopes, reviewed in Larsen *et al*. (2006). During or after the transport into the ER the peptide must bind to the MHC class I molecule (Stoltze *et al*., 2000; Zhang and Williams, 2006) before it can be transported to the cell surface through the golgi system. The most selective step in this pathway is binding of a peptide to the MHC class I molecule. In an older review, Yewdell and Bennink (1999) states that only 1 in 200 binds with an affinity strong enough to generate an immune response. This has been challenged, and it might be that up to 3% of the possible peptides bind strong enough to generate a subsequent immune response (Assarsson *et al*., 2007). In another recent work of Moutaftsi *et al*. (2006), however, it is found that of the 49 epitopes that are responsible for 95% of the total CD8+ T-cell response against a vaccinia challenge in mouse 90% binds MHC with an affinity stronger than 500 nM. In any case a peptide must go through the processes in a greater number than competing peptides to be immunodominant. The MHC is the most polymorphic gene system known. This polymorphism is a huge challenge for T-cell epitope discoveries, enhancing the need for bioinformatical analysis and resources. However, it also highly complicates immunological bioinformatics, as predictive methods for peptide MHC binding have to deal with the diverse genetic background of different populations and individuals. On a population basis, hundreds of alleles have been found for most of the HLA encoding loci (1839 in release 2.17.0 of the IMGT/HLA Database, http://www.ebi.ac.uk/imgt/hla/). In a given individual either one or two different alleles are expressed per locus depending on whether the same (in homozygous individuals) or two different (in heterozygous individuals) alleles are coded for on the two different chromosomes. The number of MHC expressing loci, however, differs highly among species. While a fully heterozygous human has six different MHC class I genes, a rhesus macaque may host up to 22 active MHC class I genes (Daza-Vamenta *et al*., 2004). Each MHC allele binds a very restricted set of peptides and the polymorphism affects the peptide binding specificity of the MHC; one MHC will recognize one part of the peptide space, whereas another MHC will recognize a different part of this space. The very large number of different MHC alleles makes reliable identification of potential epitope candidates an immense task if all alleles are to be included in the search. However, many MHC alleles share a large fraction of their peptide-binding repertoire, and it is often possible to find promiscuous peptides, which bind to a number of HLA alleles. A way of reducing the problem is to group all the different alleles into supertypes in a manner so that all the alleles within a given supertype have roughly the same peptide specificity (Hertz and Yanover, 2007; Lund *et al*., 2004; Reche and Reinherz, 2004; Sette and Sidney, 1998, 1999). This allows the search to be limited to a manageable representative set. Representing a supertype by a well-studied allele might lead to selection of epitopes that is very restricted to this allele, but not to any other alleles within the supertype. Thus another, and potentially more rational approach, would be to select a limited set of peptides restricted to as many alleles as possible. This should be within reach with new methods that directly predict epitopes that can bind to different alleles (promiscuous epitopes) (Brusic *et al*., 2002), or pan-specific approaches that can make predictions for all alleles where the sequence is known (Jojic *et al*., 2006; Nielsen *et al*., 2007a). When the peptide–MHC complex is presented on the surface of the cell, it might bind to a CD8+ T cell with a fitting T-cell receptor (TCR). If such a TCR clone exists depends on, among other factors, if the TCR–peptide complex is too similar to MHC–peptide complexes generated with peptides from the host proteome (self-peptides). This effect is called tolerance and might be broken by so-called self-epitopes, reviewed by Andersen *et al*. (2006).

B cells must be activated to produce antibodies against a given antigen, and helper T cells specific for peptides from the antigen must be activated to get a strong B-cell response. The epitope recognized by the helper T cell is usually somehow connected to the epitope that is recognized by the B cell, but the two cells do not necessarily recognize overlapping epitopes. T cells can recognize internal peptides that do not need to be a part of the surface–surface interactions with the B-cell receptor. Actually, the T-cell and the B-cell epitopes might not even come from the same protein (Janeway *et al*., 2001). The peptides recognized by the CD4+ T cells are presented by the MHC class II molecule, and peptide presentation on MHC class II molecules follow a different path than the MHC class I presentation pathway (Castellino *et al*., 1997): MHC class II molecules associate with the invariant chain (Ii) in the ER and the MHC–Ii complex accumulates in endosomal compartments. Here, Ii is degraded, while another MHC-like molecule, called HLA–DM in humans, loads the MHC class II molecules with the best available ligands originating from endocytosed antigens. The peptide–MHC class II complexes

are subsequently transported to the cell surface for presentation to T helper cells.

Immunological predictions and simulations have been demonstrated highly useful in applied immunology in general, and in vaccinology in particular. It can be used as an efficient tool to lower the experimental workload in epitope discovery for use in rational vaccine design, immunotherapeutics and development of diagnogstic tools. A number of recent publications describe in great detail the values and benefits obtained by the use of immunoinformatics and predictions in applied immunology and vaccinology (Davies and Flower, 2007; De Groot, 2006; De Groot and Moise, 2007; Korber *et al*., 2006; Lund *et al*., 2005; Petrovsky and Brusic, 2006; Tong *et al*., 2007). Here, we will not engage in this discussion, but rather limit ourselves to describing the available methods for making such predictions, and deliver some of the background information needed to be able to choose the appropriate method for a given task.

## 1.2 Prediction methods

A large variety of machine-learning techniques are commonly used in the field of immunological bioinformatics ranging from the conventional techniques of position-specific scoring matrices (PSSMs) (Altschul *et al*., 1997), Gibbs sampling (Lawrence *et al*., 1993; Nielsen *et al*., 2004), artificial neural networks (ANNs) described in Baldi and Brunak (2001), hidden Markov models (HMMs) explained in Hughey and Krogh (1996), and support vector machines (SVMs) described in Cortes and Vapnik (1995), to more exotic methods like ant colonies (Karpenko *et al*., 2005) and other motif search algorithms (Bui *et al*., 2005; Chang *et al*., 2006; Murugan and Dai, 2005). ANNs and SVMs and are ideally suited to recognize non-linear patterns, which are believed to contribute to, for instance, peptide–HLA-I interactions (Adams and Koziol, 1995; Brusic *et al*., 1994; Buus *et al*., 2003; Gulukota *et al*., 1997; Nielsen *et al*., 2003). In an ANN, information is trained and distributed into a computer network with an input layer, hidden layers and an output layer all connected in a given structure through weighted connections (Baldi and Brunak, 2001). In a PSSM on the other hand, all positions in the motif are assumed to contribute in an independent manner, and the likelihood for matching a motif is calculated as a sum of individual matrix scores. The Gibbs sampler method is a particular implementation of the PSSM search algorithm, where the optimal PSSM is determined by a search for a sequence alignment that provides maximal information content for a given motif length. Conventionally PSSMs are log-odds matrices (Altschul *et al*., 1997), where the weight matrix elements are estimated from the logarithm of the ratio of the observed frequency of a given amino acid to the background frequency of that amino acid. However, many other techniques including the stabilization matrix method (SMM) (Peters and Sette, 2005), and evolutionary algorithm (Brusic *et al*., 1998) exist to construct a PSSM. The PSSMs might also be coupled with other information available to compensate for lack of data (Lundegaard *et al*., 2004). Finally, HMMs have been used in the field of immunological bioinformatics. These are well suited to characterized biological motifs with an inherent structural composition, and have been used in the field of immunology to predict for instance peptide binding to MHC class I (Mamitsuka, 1998) and class II (Noguchi *et al*., 2002) molecules. Beside machine-learning techniques, also (empirical) molecular force field modeling techniques (Logean *et al*., 2001) and 3D Quantitative Structure–Activity Relationship (3D-QSAR) (Doytchinova and Flower, 2002; Zhihua *et al*., 2004) analysis have been used to predict features of the immune system.

## 1.3 Performance measures and validation

As an evaluation of the general quality of a prediction method a measure describing this quality is needed. However, no single measure can capture all qualities of a prediction, and not all types of data and predictions can be reasonably described by the same measure. So to be able to compare different systems, it is often needed to present several measures of quality. Most measures need the data to be classified into two groups, i.e. positives and negatives. The number of classified (experimentally measured) positives is often designated as actual positives (AP), and the number of negatives, actual negatives (AN), the number of predicted positives (PP), predicted negatives (PN), truly predicted positives (TP), falsely predicted positives (FP), truly predicted negatives (TN), and falsely predicted negatives (FN). Some of the most often used measures are briefly described here. The equations for the mentioned measures are given at the end of the section.

The fraction correct predicted (FCP) is the fraction of the total predictions that falls into the correct group. This measure is intuitively easily captured, but has the weakness that if a large fraction of the total evaluation data falls into a single group one will get high performance by just blindly predicting most or even everything to belong to this category.

The positive predicted value (PPV) is the fraction of the positive predictions that actually falls into the positive class.

The sensitivity is the fraction of the AP that is predicted as positives using a given threshold.

The specificity is the fraction of the AN that is predicted as negatives.

The three latter measures are also easily grasped, however they are all dependent on the chosen prediction cutoff classifying the data into positive and negative predictions. A high sensitivity can be obtained by setting your prediction cutoff so that most of your evaluation data will fall into the positive group, but this will then be at the expense of the specificity and the PPV. Which cutoff to use is determined by the purpose of the prediction, i.e. how many verified epitopes is needed versus the resources available for experimental validation.

A plot of the sensitivity against the false positive rate (1-specificity) is called a receiver operating characteristic (ROC) curve (Swets, 1988). Such a plot can be a help to set the best prediction cutoff. One of the best ways of measuring the predictive power of a method is to calculate the area under the ROC curve (AUC) since this is a threshold-independent measure. Another robust measure is the Pearson correlation coefficient (PCC), which is a measure of how well the prediction scores correlate with the actual value on a linear scale.

In situations where the correlation is not necessarily linear, the Spearman's rank correlation coefficient (SRC) is more appropriate. In this measure each prediction is ranked on the basis of the prediction score and the PCC is calculated on the basis of this rank rather than the prediction score. The SRC, like the AUC, is a threshold-independent measure of how well the predictor ranks the data when compared with the actual ranking.

When comparing different methods, the threshold-independent measures are to be preferred. Otherwise a threshold has to be set under the same assumptions for all predictors. As an example one can estimate the specificity for each predictor by setting the threshold for the given predictor to a value where the sensitivity will be 0.5 (i.e. half of the total available positives is over the threshold), or estimate the sensitivity at a threshold where the specificity will be 0.8 (i.e. 80% of the AN are predicted as negatives).

The choice of an evaluation set is also absolutely crucial and several considerations must be taken. A large and diverse dataset is to be preferred to avoid any biases in prediction space. Extreme care should also be taken to ensure that none of the predictors have been trained on the data used for evaluation even though that might not always be possible. To make the evaluation as broad as possible cross-validation is often used, i.e. the method is trained on a large part of the available data and a smaller part is left out for evaluation. This is done until all data has been included in the evaluation set and in this way it is possible to estimate the performance on the complete dataset. Caution has to be taken, however, that the part used for training is not too similar to the evaluation part, as this will lead to an overestimation of the performance due to over-training. This is especially true when using the leave-one-out version of cross-validation where everything except one data point is used for training, and the evaluation is then performed on the ensemble of the left out data points. Equations are as follows:

$$\begin{aligned}
\mathrm{FCP} &= (\mathrm{TP}+\mathrm{TN})/(\mathrm{AP}+\mathrm{AN}) \\
\mathrm{PPV} &= \mathrm{TP}/\mathrm{PP} \\
\text{Sensitivity} &= \mathrm{TP}/\mathrm{AP} \\
\text{Specificity} &= \mathrm{TN}/\mathrm{AN} \\
PCC &= \frac{\sum_i (a_i-\bar{a})(p_i-\bar{p})}{\sqrt{\sum_i (a_i-\bar{a})^2}\sqrt{\sum_i (p_i-\bar{p})^2}}
\end{aligned}$$

## 2 CURRENT PREDICTION ALGORITHMS

The state-of-the-art class I T-cell epitope prediction methods are today of a quality that makes it highly useful as an initial filtering technique in epitope discovery. Studies have demonstrated how it is possible to rapidly identify and verify MHC binders from upcoming possible threats such as the SARS virus (Sylvester-Hvid *et al*., 2004) with high reliability, and take such predictions a step further and validate the immunogenecity of peptides with limited efforts, as has been shown with the influenza A virus (Wang *et al*., 2007). It is also possible to identify the vast majority of the relevant epitopes in a rather complex organism as the vaccinia virus using class I MHC binding predictions and only have to test a very minor fraction of the possible peptides in the virus proteome (Moutaftsi *et al*., 2006).

MHC class II predictions can be made fairly reliable for certain alleles, and a number of helper epitopes have been identified by the help of bioinformatical approaches (Consogno *et al*., 2003).

B-cell epitopes are still the most complicated task. However, some consistency between predicted and verified epitopes is starting to emerge using the newest prediction methods (Dahlback *et al*., 2006).

In the following, we describe some of the best-performing prediction methods within each area.

### 2.1 B-cell epitope predictions

B-cell epitope prediction is a highly challenging field due to the fact that the vast majority of antibodies raised against a specific protein interact with discontinuous fragments (van Regenmortel, 1996). The prediction of continuous, or linear, epitopes, however, is a somewhat simpler problem, and may be still useful for synthetic vaccines or as diagnostic tools (Regenmortel and Muller, 1999). Moreover, the determination of continuous epitopes can be integrated into determination of discontinuous epitopes, as these often contain linear stretches (Hopp, 1994).

In the early 1980s, Hopp and Woods (Hopp and Woods, 1981, 1983) developed the first linear epitope prediction method. This method takes the assumption that the regions of proteins that have a high degree of exposure to solvent contain the antigenic determinants. According to the hydrophilicity scale generated by Levitt (1976), Hopp and Woods (1981) assigned the hydrophilicity propensity to each amino acid in a sequence and looked at groups of six residues. This gave promising results and a number of methods have since been developed with the aim of predicting linear epitopes using a combination of different amino acid propensities (Alix, 1999; Debelle *et al*., 1992; Jameson and Wolf, 1988; Maksyutov and Zagrebelnaya, 1993; Odorico and Pellequer, 2003; Parker *et al*., 1986). In 1993, Pellequer *et al*. (1993) proposed an evaluation set containing 85 continuous epitopes in 14 proteins and found that the method based on turn propensity (i.e. the propensity of an amino acid to occur within a turn structure) had the highest sensitivity using this set. Seventy percent of the residues predicted to be in epitopes by this method were actually part of epitopes. The sensitivity for methods based on other propensities was in the range of 36–61% (Pellequer *et al*., 1991). Analyzing the epitope regions in the Pellequer dataset reveals that almost all the hydrophobic amino acids are under-represented, supporting the assumption that linear B-cell epitopes will occur in hydrophilic regions of the proteins.

An extensive study of linear B-cell epitope prediction methods was published by Blythe and Flower (2005). To test how well peaks in single amino acid scale propensity profiles are (significantly) associated with known linear epitope locations, 484 amino acid propensities from the AAindex database (http://www.genome.ad.jp) (Kawashima and Kanehisa, 2000) were used. As test set they used 50 epitope-mapped proteins defined by polyclonal antibodies, which were the best non-

redundant test set available. Blythe and Flower (2005) found, however, that even the predictions based on the most accurate amino acid scales were only marginally better than random, suggesting that more sophisticated approaches is needed to predict the linear epitopes. BepiPred (Larsen *et al*., 2006), an algorithm that combines scores from the Parker hydrophilicity scale (Parker *et al*., 1986) and a PSSM trained on linear epitopes, shows a small, but significant, increase in AUC over earlier scale-based methods. The sequence parametrizer algorithm (Sollner, 2006; Sollner and Mayer, 2006), along with its associated machine-learning methods uses the common single amino acid propensity scales, but also incorporates neighborhood parameters reflecting the probability that a given stretch of amino acids exists within a predefined proximity of a specific amino acid residue. Training and testing on epitope sequences pulled from a high-quality proprietary database, as well as several publicly accessible databases, yields a degree of accuracy that is greatly increased over single-parameter methods.

Different experimental techniques can be used to define conformational epitopes. Probably the most accurate, and easily defined is using the solved structures of antibody–antigen complexes (Fleury *et al*., 2000; Mirza *et al*., 2000). The amount of this kind of data is unfortunately still scarce, compared to linear epitopes. Furthermore, very few antigens have been studied in a way where all possible epitopes on a given antigen has been identified. Unidentified epitopes within the dataset will lower the apparent performance of an accurate prediction method by increasing the apparent false positive rate.

The simplest way to predict the possible epitopes in a protein of known 3D structure is to use the knowledge of surface accessibility (Novotny *et al*., 1986; Thornton *et al*., 1986). Two newer methods using protein structure and surface exposure for prediction of B-cell epitopes have been developed. The CEP method (Kulkarni-Kale *et al*., 2005) calculates the relative accessible surface area for each residue in the structure. Then it is determined which parts of the protein that are exposed enough to be antigenic determinants. Regions that are distant in the primary sequence, but close in three-dimensional space are considered as one epitope. The tool was tested on a dataset of 63 antigen–antibody complexes and the algorithm correctly identified 76% of the epitope residues. DiscoTope (Haste *et al*., 2006) uses a combination of amino acid statistics, spatial information and surface exposure. It is trained on a compiled dataset of discontinuous epitopes from 76 X-ray structures of antibody–antigen protein complexes. This method outperforms methods that predict linear epitopes. Recently a workshop was held on the subject of B-cell epitope predictions attended by a broad range of the current method developers. The workshop resulted in a published review containing conclusions on the present common ground, and suggestions for the future especially concerning coordination and evaluation (Greenbaum *et al*., 2007).

Different ways of measuring the accuracy of B-cell epitope predictions have been suggested (Hopp, 1994; van Regenmortel and Pellequer, 1994). Pellequer suggested using the specificity as a measure of accuracy, while Hopp suggested using the PPV, but, as described earlier, neither measure will alone give a good description of the performance. In accordance to this the recent workshop concluded that the AUC measure is to be preferred (Greenbaum *et al*., 2007). Another issue is whether to make the statistics on a per-residue or on a per-epitope basis. However, as the latter have the additional complications of defining how much of an epitope that must be included in a prediction to be considered correct, and how much extra included residues is allowed, the per residue measure is to be preferred.

Epitope mapping can be performed experimentally by other methods than structure determination, e.g. by phage display (Jesaitis *et al*., 1999; Smith and Petrenko, 1997). The low sequence similarity between the mimotope [i.e. a macromolecule, often a peptide, which mimics the structure of an epitope, (Meloen *et al*., 2000)] identified through phage display and the antigen complicates the mapping back onto the native structure of the antigen. A number of methods have been developed to facilitate this (Batori *et al*., 2006; Enshell-Seijffers *et al*., 2003; Halperin *et al*., 2003; Huang *et al*., 2006; Moreau *et al*., 2006; Mumey *et al*., 2003; Schreiber *et al*., 2005; Tarnovitski *et al*., 2006). However, these are to be considered as interpreters of experimental data rather than predictors, which are the main focus of this review.

## 2.2 MHC binding

A number of methods for predicting the binding of peptides to MHC molecules have been developed (Schirle *et al*., 2001) since the first motif methods were presented (Rothbard and Taylor, 1988; Sette *et al*., 1989). The majority of peptides binding to MHC class I molecules have a length of 8–10 amino acids. Position 2 and the C-terminal position have turned out generally to be very important for the binding to most class I MHCs and these positions are referred to as anchor positions (Rammensee *et al*., 1999). For some alleles, the binding motifs further have auxiliary anchor positions. Peptides binding to the human HLA-A*0101 allele thus have positions 2, 3 and 9 as anchors (Kondo *et al*., 1997; Kubo *et al*., 1994; Rammensee *et al*., 1999). The importance of anchor positions for peptide binding and the allele-specific amino acid preference at the anchor positions was first described by Falk *et al*., 1990. The discovery of such allele-specific motifs led to the development of the first reasonable accurate algorithms (Pamer *et al*., 1991; Rotzschke *et al*., 1991). In these prediction tools, it is assumed that the amino acids at each position along the peptide sequence contribute a given binding energy, which can independently be added up to yield the overall binding energy of the peptide (Meister *et al*., 1995; Parker *et al*., 1994; Stryhn *et al*., 1996). Similar types of approaches are used by the EpiMatrix method (Schafer *et al*., 1998), the BIMAS method (Parker *et al*., 1994), the SYFPEITHI method (Rammensee *et al*., 1999), the RANKPEP method (Reche *et al*., 2002) and the Gibbs sampler method (Nielsen *et al*., 2004). Several of these matrix methods use an approach in the development where the method is build using exclusively positive examples defined after certain criteria, like eluted peptides and interferon gamma response data. This data can be used in training as well as affinity binding data defining binding stronger than a certain threshold (usually 500 nM). Other matrix methods, like the SMM method, aim at predicting an actual affinity and thus

use exclusively affinity data. As described earlier, matrix-based methods cannot take correlated effects into account, i.e. where the contribution to the binding affinity by a given amino acid at one position is influenced by amino acids at other positions in the peptide. Higher order methods like ANNs and SVMs, on the other hand, are ideally suited to take such correlations into account. These methods can be trained with data either in the format of binder/non-binder classification, or as real affinity data. Some of the recent methods combine the two types of data and prediction methods, either by averaging over predictions made by either (Bhasin and Raghava, 2007), or by feeding the predictions from the positive data-trained PSSMs to ANNs together with sequence/affinity data (Nielsen *et al*., 2003). A study by Yu *et al*. (2002) clearly shows the influence of having a large dataset on the performance of the resulting method. However, including knowledge of important positions reduce the need for data significantly (Lundegaard *et al*., 2004).

Several prediction methods have been made publicly available, and when selecting between these several cautions should be taken. The published performance, and how it is evaluated should be examined, but it is also very important that the method is able to generate predictions for the actual allele of interest. A major study comparing the predictive performance of a large part of the available methods was recently performed by Peters *et al*. (2006) showing that in general the SMM and the ANN methods (Table 1) perform the best, even when taken into account the number of training data for each method. The cross-validated performance of these methods for several human and mouse MHC class I alleles was compared with the best performing other method available as web tool. The full results of this work are listed in Supplementary Table 1. The tools and URLs are listed in Table 1. It should be mentioned, however, that tools known to be trained on a significant part of the test set were excluded from this comparison. To achieve

**Table 1.** URLs for a selected subset of the methods in Peters *et al*. (2006)

| Name | URL |
|---|---|
| IEDB[a] | http://tools.immuneepitope.org/analyze/html/mhc_binding.html |
| NetMHC[b] | http://cbs.dtu.dk/services/NetMHC |
| BIMAS | http://thr.cit.nih.gov/cgi-bin/molbio/ken_parker_comboform |
| hla_a2_smm | http://zlab.bu.edu/SMM-cgi/peptide1.cgi |
| hlaligand | http://hlaligand.ouhsc.edu/prediction.htm |
| libscore | http://hypernig.nig.ac.jp/cgi-bin/Lib-score/request.rb |
| mhcpred | http://www.jenner.ac.uk/MHCPred/ |
| multipredann | http://research.i2r.a-star.edu.sg/multipred/HTML/predict.html |
| pepdist | http://www.pepdist.cs.huji.ac.il/ |
| predbalbc | http://antigen.i2r.a-star.edu.sg/predBalbc/ |
| rankpep | http://mif.dfci.harvard.edu/Tools/rankpep.html |
| svmhc | http://www.sbc.su.se/svmhc/new.cgi |
| syfpeithi | http://www.syfpeithi.de/ |

[a]The SMM, ARB, and ANN methods from Peters *et al*. (2006).
[b]Updated version of the ANN method from Peters *et al*. (2006).

binding predictions for an allele with uncharacterized specificity, the supertype concept (Sette and Sidney, 1998) can be used for the limited number of alleles with well-defined supertype relationships (Lund *et al*., 2005). Note, however, that predictions with methods predicting the specific allele is most often to be preferred, as the accuracy of these will be better (Nielsen *et al*., 2007a).

In general, HLA-I binding predictions depend on sufficient experimental data being available for the exact HLA-I molecule in question. Unfortunately, <10% of the 1500 registered HLA-I proteins (Lefranc, 2005) have been examined experimentally, and <5% have been characterized with more than 50 examples of peptide binders (Rammensee *et al*., 1999; Sette *et al*., 2005).

Several groups have suggested prediction strategies to span these 'uncharacterized' regions of the HLA diversity (Brusic *et al*., 2002; Jojic *et al*., 2006; Nielsen *et al*., 2007; Zhu *et al*., 2006). In different forms, all these methods exploit both peptide and primary HLA sequence as input information for training, aiming at simultaneously incorporating all HLA specificities. In a recent paper (Nielsen *et al*., 2007a), it is successfully demonstrated that such an approach can, to a very high degree, accurately characterize the binding motif for previously untested HLA-I molecules.

Unlike the MHC class I molecules, the binding cleft of MHC class II molecules is open-ended, which allows for the bound peptide to have significant overhangs in both ends. As a result MHC class II binding peptides have a broader length distribution even though the part of the binding peptide that interacts with the MHC (the binding core) still includes only 9 amino acid residues. This complicate binding predictions as identification of the correct alignment of the binding core is a crucial part of identifying the MHC class II binding motif (Nielsen *et al*., 2004). The MHC class II binding motifs have relatively weak and often degenerate sequence signals. While some alleles like HLA-DRB1*0405 show a strong preference for certain amino acids at the anchor positions, other alleles like HLA-DRB1*0401 allow basically all amino acids at all positions (Rammensee *et al*., 1999). However, there are other issues affecting the predictive performance of most MHC class II binding prediction methods. The majority of these methods take as a fundamental assumption that the peptide–MHC binding affinity is determined solely from the nine amino acids in binding core motif. This is clearly a large oversimplification since it is known that peptide flanking residues (PFR) on both sides of the binding core may contribute to the binding affinity and stability (Godkin *et al*., 2001). Some methods for MHC class II binding have attempted to include PFRs indirectly, in terms of the peptide length, in the prediction of binding affinities (Chang *et al*., 2006). Recently, Nielsen *et al*. (2007b) published a method for MHC class II prediction that directly include PFRs and demonstrated that these PFRs improves the prediction accuracy. Most of the methods for MHC class II binding predictions have been trained and evaluated on very limited datasets covering only a single or a few different MHC class II alleles, making it very difficult to compare the different performance values and generality of the methods. Nielsen *et al*. (2007b) have made available a large-scale benchmark set-up for evaluating MHC class II peptide binding affinity prediction algorithms. The benchmark covers

14 HLA-DR (human MHC) and three mouse H2-IA alleles, and consists of peptide/IC50 affinity data downloaded from the publicly available IEDB database (Peters *et al*., 2005), and could set the start for large-scale unbiased evaluations of novel methods for MHC class II prediction.

### 2.3 Processing

Successful prediction of the proteasome cleavage site specificity should provide valuable additional information useful in the design of treatments based on CTL responses. However, the complexity of proteasomal enzymatic specificity complicates such predictions. The proteasome have a highly stochastic element, exemplified by the observation that only ~80% of the cleavage sites observed in one *in vitro* experiment can be verified in a second identical experiment (Hansjörg Schild, personal communication). It is thus expected that the accuracy for prediction of proteasomal activity will be relatively low when compared to that of methods for MHC peptide binding.

FragPredict, which is publicly available as a part of MAPPP service (http://www.mpiibberlin.mpg.de/MAPPP/), combines proteasomal cleavage predictions with MHC- and TAP-binding predictions. FragPredict consists of two algorithms. The first algorithm uses a statistical analysis of cleavage-enhancing and -inhibiting amino acid motifs to predict potential proteasomal cleavage sites (Holzhutter *et al*., 1999). The second algorithm, which uses the results of the first algorithm as an input, predicts which fragments are most likely to be generated. This model takes the time-dependent degradation into account based on a kinetic model of the 20S proteasome (Holzhutter and Kloetzel, 2000). At the moment, FragPredict is the only method that can predict fragments, instead of only possible cleavage sites.

PAProC (http://www.paproc.de) is a prediction method for cleavages by human as well as wild type and mutant yeast proteasomes. The influences of different amino acids at different positions are determined by using a stochastic hillclimbing algorithm (Kuttler *et al*., 2000) based on the experimentally *in vitro* verified cleavage and non-cleavage sites (Nussbaum *et al*., 2001). Both the FragPredict and PAProC methods make use of the limited *in vitro* proteasomal digest data available. FragPredict is a linear method, and it may not capture the non-linear features of the specificity of the proteasome. The NetChop (Kesmir *et al*., 2002) method tries to address these two issues. The prediction system is a multilayered ANN and uses naturally processed MHC class I ligands to predict proteasomal cleavage. Since some of these ligands are generated by the immunoproteasome, and some by the constitutive proteasome, such a method should predict the combined specificity of both forms of proteasomes. In 2003, NetChop-2.0 were evaluated to be the best-performing predictor on an independent evaluation set (Saxová *et al*., 2003). Pcleavage is another web accessible proteasomal cleavage predictor, which is SVM based and have a published performance comparable to NetChop-2.0 (Bhasin and Raghava, 2005). An update of the NetChop method [NetChop-3.0, Nielsen *et al*. (2005)] consists of a combination of several ANNs, each trained using a different sequence-encoding scheme of the data. NetChop 3.0 has an increase in the prediction sensitivity as compared to NetChop 2.0, without lowering the specificity, and is thus probably the current best predictor of proteasomal cleavage. Tenzer *et al*. (2004) have published a weight matrix based method for prediction of both constitutive- and immunoproteasomal cleavage specificity. Both matrices are trained on *in vitro* digest data.

Relatively few methods have been developed to predict the specificity of TAP. Daniel *et al*. (1998) have developed ANNs using peptide 9mers for which TAP affinity was determined experimentally. Surprisingly, they found that some MHC alleles have ligands with very low TAP affinities, e.g. HLA-A2. However, it has been shown that TAP ligands can be trimmed in ER before binding to MHC molecules (Fruci *et al*., 2001), i.e. a TAP ligand might be an epitope precursor and thus does not need to be 9 amino acids long. HLA-A2 might easily have precursors of its optimal ligands, which are also good TAP binders. Peters *et al*. (2003) used an SMM to predict TAP affinity of peptides. This method has the advantage of not being bound to only 9mers but can also be used for longer peptides. The method assumes that only the first three positions in the N-terminal and the last position at the C-terminal influences the TAP binding. The method is very well evaluated and the accuracy is high. The significance of TAP binding in the epitope presentation pathway is much lower than the MHC binding (see later) and the AUC value when this method is used alone as an epitope predictor of 0.79 is thus significantly lower than most MHC-binding prediction methods. Two methods were published in 2004. Bhasin and Raghava (2004) published a method for which they do only compare to the method of Daniel *et al*. (1998) and it is not determined how it performs compared to the Peters' method. The method of Doytchinova *et al*. (2004) is evaluated by comparing the resulting method (matrix) with other matrices. From such a comparison it can only be concluded that this method is closer to Peters' model than to the model of Bhasin and Raghava (2004) but not how it actually performs. Recently a new TAP predictor, PredTAP, have been published (Zhang *et al*., 2006). This method does not have an AUC value for the methods performance in epitope prediction making a direct comparison to other models impossible. With increasing numbers of TAP ligands available on the internet (e.g. Jen-Pep database, http://www.jenner.ac.uk) (Blythe *et al*., 2002), it will likely soon be possible to obtain more accurate TAP predictions.

With respect to TAP-independent transport and cleavage of peptides, the most established model is especially connected to the most abundant HLA supertype (A2) and is related to the signal peptides and the processing of such (Larsen *et al*., 2006). Prediction of potential signal peptides that can be transported by Sec61 can be made with tools for prediction of signal peptides, and some of these will also predict the signal peptidase cleavage site (Bendtsen *et al*., 2004; Kall *et al*., 2004; Zhang and Henzel, 2004), but the value in the context of CD8+ T-cell epitope predictions remains to be elucidated.

The TCRs are generated by highly stochastic processes that secures that the TCRs in general will be able to recognize the entire probable space of MHC–peptide complexes. However, TCRs that recognize self-peptides will be eliminated so peptides that form complex with MHC are indistinguishable from

self-peptides will not be recognized. It is still not clear how close peptides must be to the self to be able to escape recognition in this way (Louzoun *et al*., 2006).

### 2.4 Integrated T-cell epitope predictions

Reliable predictions of immunogenic peptides can reduce the experimental effort needed to identify new epitopes, and though reliable predictions of the MHC binding alone can indeed be used to rank the possible epitopes very accurately, even better predictions should be possible if the other steps in the pathway were integrated in the predictions. Accordingly, many attempts have been made to predict the outcome of the steps involved in antigen presentation, MAPP (Hakenberg *et al*., 2003), NetCTL (Larsen *et al*., 2005), MHCpathway (Tenzer *et al*., 2005), epiJen (Doytchinova *et al*., 2006) and WAPP (Donnes and Kohlbacher, 2005). All these methods attempt to predict antigen presentation by integrating peptide–MHC binding predictions with one or more of the other events involved in the antigen presentation pathway. To benchmark these, a set of verified epitopes can be used as the positive dataset. Negative examples (peptides that cannot induce an immunologic response) are hard to identify, as it is very hard to determine that a peptide will never be an epitope in any persons with a given HLA haplotype. Instead, epitopes from well-studied pathogens (e. g. HIV) are often used as the positive set, and all other peptides from the genome of the same pathogen that have never been shown to be an epitope are assumed negative as they have a very low probability of being an epitope. Running a large-scale benchmark calculation comparing the predictive performance of several publicly available MHC-I presentation prediction methods evaluated on a large set of known HIV epitopes (http://www.cbs.dtu.dk/suppl/immunology/CTL-1.2/HIV_dataset) reveals that the updated NetCTL and MHCpathway methods have the highest predictive performance with >75% if the epitopes being within the top 5% peptides with the highest prediction scores (Mette Volby Larsen, personal communication).

## 3 SIMULATING THE IMMUNE SYSTEM

Improved understanding of the immune systems, and its population-wide variation, is one of the major challenges in the next decade within biology and medicine. Many of the steps by which the immune system deal with infectious agents and disease can now successfully be modeled by computational techniques, and it is clear that the theoretical approaches will be a major player in this area, adding a systems view to the massive experimental effort being carried out at the moment. In this review, we have summarized how a number of bioinformatics tools that use genomic sequences as input to predict epitopes, have been developed over the past decade. At the same time, theoretical models have been developed that describe the dynamics of different immune-cell populations and their interactions with microbes (Borghans and de Boer, 2007; Carneiro *et al*., 2007; Davenport *et al*., 2007). These models have been used to interpret experimental findings where timing is of importance, such as the interval between administration of a vaccine and infection with the microbe that the vaccine is intended to protect against. Moreover, these dynamic models allowed for generating a quantitative picture of immune system kinetics and diversity during health and disease. The quantitative approach is necessary to understand the functioning of the immune system, which consists of many different cell types and molecules interacting in complicated regulatory pathways involving positive and negative feedback loops. Surprisingly little is known about the population dynamics, i.e. the production rates, division rates and distribution of life spans of mouse or human lymphocyte populations. As a consequence, fundamental questions like the maintenance of memory, the maintenance of a diverse naive repertoire and the role of homeostatic mechanisms, remain largely unresolved. Having so little insight in the normal lymphocyte population dynamics also hampers our understanding of immune responses during disease and immune reconstitution after therapeutic interventions such as chemotherapy, irradiation and/or bone marrow transplantation. Several areas in immunology call for a better interpretation of data by means of theoretical models. A simple PubMed search reveals that at least 10% of the recent papers in the immunological literature involve labeling experiments in which lymphocytes are labeled radioactively, with deuterium, or with dyes. However, the interpretation of such labeling data is controversial and is notoriously difficult (Boer *et al*., 2003a, b; Deenick *et al*., 2003; Gett and Hodgkin, 2000; Hellerstein, 1999; Mohri *et al*., 1998; Mohri *et al*., 2001; Revy *et al*., 2001; Ribeiro *et al*., 2002), which emphasizes the enormous demand to develop a quantitative mathematical approach to immunology. Similar examples of how difficult it is to properly interpret kinetic data come from the attempts to characterize the division history of cells from the length of the telomeres, or from the presence of autosomal DNA circles (TRECs) that are formed in the thymus (Boer and Noest, 1998; Douek *et al*., 1998; Dutilh and de Boer, 2003; Hazenberg *et al*., 2000; Hazenberg *et al*., 2003).

Integrating the dynamic (using mathematical models and computer simulations) and bioinformatics approaches clearly could lead to a better understanding of the immune responses and their role during normal, disease and reconstitution states, where both timing and sequence specificity are highly significant. Diseases that are characterized by complex interactions between the host cellular immune system and evolving pathogens such as HIV infection, or diseases where molecular similarities between self and non-self are important such as in autoimmune diseases could be investigated in such integrated models. Complex generalized cellular automata have been proposed as models of the immune system (Kohler *et al*., 2000; Seiden and Celada, 1992). These methods have now developed to a stage where it is possible successfully to simulate the outcome of cancer vaccine protocols using a mouse simulation model (Castiglione and Piccoli, 2007; Lollini *et al*., 2006; Motta *et al*., 2005; Pappalardo *et al*., 2006). In a recent paper, Rapin *et al*. (2006) outline a framework for integration of these bioinformatics and simulation approaches by developing a simple model in which HIV dynamics are correlated with genomics data. This model is the first one where, the fitness of wild-type and mutated virus is assessed by means of a sequence-dependent scoring matrix that links protein sequences to growth rates of the virus. Further

refinements of these approaches may involve increasing the spatial resolution by including different tissues and their geometry.

## ACKNOWLEDGEMENTS

This work was funded by European Commission (LSHB-CT-2003-503231, LSHB-CT-2004-012175) and National Institutes of Health (HHSNN26600400006C, HHSN266200400025C, HHSN266200400083C).

*Conflict of Interest*: none declared.

## REFERENCES

Abele,R. and Tampé,R. (2004) The ABCs of immunology: structure and function of TAP, the transporter associated with antigen processing. *Physiology*, **19**, 216–224.

Adams,H.P. and Koziol,J.A. (1995) Prediction of binding to MHC class I molecules. *J. Immunol. Methods*, **185**, 181–190.

Alix,A.J. (1999) Predictive estimation of protein linear epitopes by using the program PEOPLE. *Vaccine*, **18**, 311–314.

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Andersen,M.H. *et al.* (2006) Cytotoxic T cells. *J. invest. dermatol.*, **126**, 32–41.

Assarsson,E. *et al.* (2007) A quantitative analysis of the variables affecting the repertoire of T cell specificities recognized after vaccinia virus infection. *J. Immunol.*, **178**, 7890–7901.

Baldi,P. and Brunak,S. (2001) *Bioinformatics: The Machine Learning Approach,* 2nd edition. MIT Press, Cambridge, Mass.

Batori,V. *et al.* (2006) An in silico method using an epitope motif database for predicting the location of antigenic determinants on proteins in a structural context. *J. Mol. Recognit.*, **19**, 21–29.

Bendtsen,J.D. *et al.* (2004) Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.*, **340**, 783–795.

Bhasin,M. and Raghava,G.P. (2004) Analysis and prediction of affinity of TAP binding peptides using cascade SVM. *Protein Sci.*, **13**, 596–607.

Bhasin,M. and Raghava,G.P. (2007) A hybrid approach for predicting promiscuous MHC class I restricted T cell epitopes. *J. Biosci.*, **32**, 31–42.

Bhasin,M. and Raghava,G.P.S. (2005) Pcleavage: an SVM based method for prediction of constitutive proteasome and immunoproteasome cleavage sites in antigenic sequences. *Nucleic Acids Res.*, **33**, W202–W207.

Blythe,M.J. and Flower,D.R. (2005) Benchmarking B cell epitope prediction: underperformance of existing methods. *Protein Sci.*, **14**, 246–248.

Blythe,M.J. *et al.* (2002) JenPep: a database of quantitative functional peptide data for immunology. *Bioinformatics*, **18**, 434–439.

Boer,R.J.d. and Noest,A.J. (1998) T cell renewal rates, telomerase, and telomere length shortening. *J. Immunol.*, **160**, 5832–5837.

Boer,R.J.d. *et al.* (2003a) Different dynamics of CD4+ and CD8+ T cell responses during and after acute lymphocytic choriomeningitis virus infection. *J. Immunol.*, **171**, 3928–3935.

Boer,R.J.d. *et al.* (2003b) Estimating average cellular turnover from 5-bromo-2′-deoxyuridine (BrdU) measurements. *Proceedings*, **270**, 849–858.

Borghans,J.A.M. and de Boer,R.J. (2007) Quantification of T-cell dynamics: from telomeres to DNA labeling. *Immunol. Rev.*, **216**, 35–47.

Brusic,V. *et al.* (1994) Prediction of MHC binding peptides using artificial neural networks. In Stonier,R.J. and Yu,X.S. (ed.) *Complex Systems: Mechanism of Adaptation*. Amsterdam, IOS Press, pp. 253–260.

Brusic,V. *et al.* (1998) Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network. *Bioinformatics*, **14**, 121–130.

Brusic,V. *et al.* (2002) Prediction of promiscuous peptides that bind HLA class I molecules. *Immunol. Cell. Biol.*, **80**, 280–285.

Bui,H.H. *et al.* (2005) Automated generation and evaluation of specific MHC binding predictive tools: ARB matrix applications. *Immunogenetics*, **57**, 304–314.

Burton,D.R. (2002) Antibodies, viruses and vaccines. *Nat. Rev. Immunol.*, **2**, 706–713.

Buus,S. *et al.* (2003) Sensitive quantitative predictions of peptide-MHC binding by a 'Query by Committee' artificial neural network approach. *Tissue Antigens*, **62**, 378–384.

Carneiro,J. *et al.* (2007) When three is not a crowd: a crossregulation model of the dynamics and repertoire selection of regulatory CD4+ T cells. *Immunol. Rev.*, **216**, 48–68.

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

Castiglione,F. and Piccoli,B. (2007) Cancer immunotherapy, mathematical modeling and optimal control. *J.Theor. Biol.*, **247**, 723–732.

Chang,S.T. *et al.* (2006) Peptide length-based prediction of peptide-MHC class II binding. *Bioinformatics*, **22**, 2761–2767.

Consogno,G. *et al.* (2003) Identification of immunodominant regions among promiscuous HLA-DR-restricted CD4+ T-cell epitopes on the tumor antigen MAGE-3. *Blood*, **101**, 1038–1044.

Cortes,C. and Vapnik,V. (1995) Support-vector networks. *Mach. Learn.*, **20**, 273–297.

Dahlback,M. *et al.* (2006) Epitope mapping and topographic analysis of VAR2CSA DBL3X involved in *P. falciparum* placental sequestration. *PLoS pathog.*, **2**, e124.

Daniel,S. *et al.* (1998) Relationship between peptide selectivities of human transporters associated with antigen processing and HLA class I molecules. *J. Immunol.*, **161**, 617–624.

Davenport,M.P. *et al.* (2007) Understanding the mechanisms and limitations of immune control of HIV. *Immunol. Rev.*, **216**, 164–175.

Davies,M.N. and Flower,D.R. (2007) Harnessing bioinformatics to discover new vaccines. *Drug Discov. Today*, **12**, 389–395.

Daza-Vamenta,R. *et al.* (2004) Genetic divergence of the rhesus Macaque major histocompatibility complex. *Genome Res.*, **14**, 1501–1515.

de Groot,A.S. (2006) Immunomics: discovering new targets for vaccines and therapeutics. *Drug Discov. Today*, **11**, 203–209.

de Groot,A.S. and Moise,L. (2007) Prediction of immunogenicity for therapeutic proteins: state of the art. *Curr. Opin. Drug Discov. Devel.*, **10**, 332–340.

Debelle,L. *et al.* (1992) Predictions of the secondary structure and antigenicity of human and bovine tropoelastins. *Eur. Biophys. J.*, **21**, 321–329.

Deenick,E.K. *et al.* (2003) Stochastic model of T cell proliferation: a calculus revealing IL-2 regulation of precursor frequencies, cell cycle time, and survival. *J. Immunol.*, **170**, 4963–4972.

Donnes,P. and Kohlbacher,O. (2005) Integrated modeling of the major events in the MHC class I antigen processing pathway. *Protein Sci.*, **14**, 2132–2140.

Douek,D.C. *et al.* (1998) Changes in thymic function with age and during the treatment of HIV infection. *Nature*, **396**, 690–695.

Doytchinova,I. *et al.* (2004) Transporter associated with antigen processing preselection of peptides binding to the MHC: a bioinformatic evaluation. *J. Immunol.*, **173**, 6813–6819.

Doytchinova,I.A. and Flower,D.R. (2002) Physicochemical explanation of peptide binding to HLA-A*0201 major histocompatibility complex: a three-dimensional quantitative structure-activity relationship study. *Proteins*, **48**, 505–518.

Doytchinova,I.A. *et al.* (2006) EpiJen: a server for multistep T cell epitope prediction. *BMC Bioinformatics*, **7**, 131.

Dutilh,B.E. and de Boer,R.J. (2003) Decline in excision circles requires homeostatic renewal or homeostatic death of naive T cells. *J. Theor. Biol.*, **224**, 351–358.

Enshell-Seijffers,D. *et al.* (2003) The mapping and reconstitution of a conformational discontinuous B-cell epitope of HIV-1. *J. Mol. Biol.*, **334**, 87–101.

Falk,K. *et al.* (1990) Cellular peptide composition governed by major histocompatibility complex class I molecules. *Nature*, **348**, 248–251.

Fleury,D. *et al.* (2000) Structural evidence for recognition of a single epitope by two distinct antibodies. *Proteins*, **40**, 572–578.

Fruci,D. *et al.* (2001) Efficient MHC class I-independent amino-terminal trimming of epitope precursor peptides in the endoplasmic reticulum. *Immunity*, **15**, 467–476.

Gett,A.V. and Hodgkin,P.D. (2000) A cellular calculus for signal integration by T cells. *Nature immunol.*, **1**, 239–244.

Godkin,A.J. *et al.* (2001) Naturally processed HLA class II peptides reveal highly conserved immunogenic flanking region sequence preferences that reflect antigen processing rather than peptide-MHC interactions. *J. Immunol.*, **166**, 6720–6727.

Greenbaum,J.A. *et al.* (2007) Towards a consensus on datasets and evaluation metrics for developing B-cell epitope prediction tools. *J. Mol. Recognit.*, **20**, 75–82.

Gulukota,K. *et al.* (1997) Two complementary methods for predicting peptides binding major histocompatibility complex molecules. *J. Mol. Biol.*, **267**, 1258–1267.

Hakenberg,J. *et al.* (2003) MAPPP: MHC class I antigenic peptide processing prediction. *Appl. Bioinformatics*, **2**, 155–158.

Halperin,I. *et al.* (2003) Sitelight: binding-site prediction using phage display libraries. *Protein Sci.*, **12**, 1344–1359.

Haste,A.P. *et al.* (2006) Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. *Protein Sci.*, **15**, 2558–2567.

Hazenberg,M.D. *et al.* (2000) Increased cell division but not thymic dysfunction rapidly affects the T-cell receptor excision circle content of the naive T cell population in HIV-1 infection. *Nat. med.*, **6**, 1036–1042.

Hazenberg,M.D. *et al.* (2003) Thymic output: a bad TREC record. *Nature immunol.*, **4**, 97–99.

Hellerstein,M.K. (1999) Measurement of T-cell kinetics: recent methodologic advances. *Immunol. Today*, **20**, 438–441.

Hertz,T. and Yanover,C. (2007) Identifying HLA supertypes by learning distance functions. *Bioinformatics*, **23**, e148–155.

Holzhutter,H.G. and Kloetzel,P.M. (2000) A kinetic model of vertebrate 20{S} proteasome accounting for the generation of major proteolytic fragments from oligomeric peptide substrates. *Biophys. J.*, **79**, 1196–1205.

Holzhutter,H.G. *et al.* (1999) A theoretical approach towards the identification of cleavage-determining amino acid motifs of the 20 S proteasome. *J. Mol. Biol.*, **286**, 1251–1265.

Hopp,T.P. (1994) Different views of protein antigenicity. *Pept. Res.*, **7**, 229–231.

Hopp,T.P. and Woods,K.R. (1981) Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl. Acad. Sci. USA*, **78**, 3824–3828.

Hopp,T.P. and Woods,K.R. (1983) A computer program for predicting protein antigenic determinants. *Mol. immunol.*, **20**, 483–489.

Huang,J. *et al.* (2006) MIMOX: a web tool for phage display based epitope mapping. *BMC Bioinformatics*, **7**, 451.

Hughey,R. and Krogh,A. (1996) Hidden Markov models for sequence analysis: extension and analysis of the basic method. *Comput. Appl. Biosci.*, **12**, 95–107.

Jameson,B.A. and Wolf,H. (1988) The antigenic index: a novel algorithm for predicting antigenic determinants. *Comput. Appl. Biosci.*, **4**, 181–186.

Janeway,C.A. *et al.* (2001) *Immunobiology: The Immune System in Health and Disease*. Garland Publications, New York, London.

Jesaitis,A.J. *et al.* (1999) Actin surface structure revealed by antibody imprints: evaluation of phage-display analysis of anti-actin antibodies. *Protein Sci.*, **8**, 760–770.

Jojic,N. *et al.* (2006) Learning MHC I-peptide binding. *Bioinformatics*, **22**, e227–235.

Kall,L. *et al.* (2004) A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.*, **338**, 1027–1036.

Karpenko,O. *et al.* (2005) Prediction of MHC class II binders using the ant colony search strategy. *Artif. Intell. Med.*, **35**, 147–156.

Kawashima,S. and Kanehisa,M. (2000) AAindex: amino acid index database. *Nucleic Acids Res.*, **28**, 374.

Kesmir,C. *et al.* (2002) Prediction of proteasome cleavage motifs by neural networks. *Protein Eng.*, **15**, 287–296.

Kohler,B. *et al.* (2000) A systematic approach to vaccine complexity using an automaton model of the cellular and humoral immune system. I. viral characteristics and polarized responses. *Vaccine*, **19**, 862–876.

Kondo,A. *et al.* (1997) Two distinct HLA-A*0101-specific submotifs illustrate alternative peptide binding modes. *Immunogenetics*, **45**, 249–258.

Korber,B. *et al.* (2006) Immunoinformatics comes of age. *PLoS Comput. Biol.*, **2**, e71.

Kubo,R.T. *et al.* (1994) Definition of specific peptide motifs for four major HLA-A alleles. *J. Immunol.*, **152**, 3913–3924.

Kulkarni-Kale,U. *et al.* (2005) CEP: a conformational epitope prediction server. *Nucleic Acids Res.*, **33**, W168–171.

Kuttler,C. *et al.* (2000) An algorithm for the prediction of proteasomal cleavages. *J. Mol. Biol.*, **298**, 417–429.

Larsen,J.E. *et al.* (2006) Improved method for predicting linear B-cell epitopes. *Immunome Res.*, **2**, 2.

Larsen,M.V. *et al.* (2005) 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.*, **35**, 2295–2303.

Larsen,M.V. *et al.* (2006) TAP-independent MHC class I presentation. *Curr. Immunol. Rev.*, **2**, 233–245.

Lawrence,C.E. *et al.* (1993) Detecting sutble sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.

Lefranc,M.P. (2005) IMGT, the international ImMunoGeneTics information system(R): a standardized approach for immunogenetics and immunoinformatics. *Immunome Res.*, **1**, 3.

Levitt,M. (1976) A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.*, **104**, 59–107.

Li,Z. *et al.* (2004) The generation of antibody diversity through somatic hypermutation and class switch recombination. *Genes Dev.*, **18**, 1–11.

Logean,A. *et al.* (2001) Customized versus universal scoring functions: application to class I MHC-peptide binding free energy predictions. *Bioorg. Med. Chem. Lett.*, **11**, 675–679.

Lollini,P.L. *et al.* (2006) Discovery of cancer vaccination protocols with a genetic algorithm driving an agent based simulator. *BMC Bioinformatics*, **7**, 352.

Loureiroa,J. and Ploegha,H.L. (2006) Antigen presentation and the ubiquitin-proteasome system in host–pathogen interactions. *Adv. Immunol.*, **92**, 225–305.

Louzoun,Y. *et al.* (2006) T-cell epitope repertoire as predicted from human and viral genomes. *Mol. Immunol.*, **43**, 559–569.

Lund,O. *et al.* (2004) Definition of supertypes for HLA molecules using clustering of specificity matrices. *Immunogenetics*, **55**, 797–810.

Lund,O. *et al.* (2005) *Immunological Bioinformatics*. MIT Press, Cambridge, MA.

Lundegaard,C. *et al.* (2004) MHC class I epitope binding prediction trained on small data sets. In *Artificial Immune Systems, Proceedings*. Springer, pp. 217–225.

Maksyutov,A.Z. and Zagrebelnaya,E.S. (1993) ADEPT: a computer program for prediction of protein antigenic determinants. *Comput. Appl. Biosci.*, **9**, 291–297.

Mamitsuka,H. (1998) Predicting peptides that bind to MHC molecules using supervised learning of hidden Markov models. *Proteins*, **33**, 460–474.

Meister,G.E. *et al.* (1995) Two novel T cell epitope prediction algorithms based on MHC-binding motifs; comparison of predicted and published epitopes from Mycobacterium tuberculosis and HIV protein sequences. *Vaccine*, **13**, 581–591.

Meloen,R.H. *et al.* (2000) Mimotopes: realization of an unlikely concept. *J. Mol. Recognit.*, **13**, 352–359.

Mirza,O. *et al.* (2000) Dominant epitopes and allergic cross-reactivity: complex formation between a Fab fragment of a monoclonal murine IgG antibody and the major allergen from birch pollen Bet v 1. *J. Immunol.*, **165**, 331–338.

Mohri,H. *et al.* (1998) Rapid turnover of T lymphocytes in SIV-infected rhesus macaques. *Science*, **279**, 1223–1227.

Mohri,H. *et al.* (2001) Increased turnover of T lymphocytes in HIV-1 infection and its reduction by antiretroviral therapy. *J. Exp. Med.*, **194**, 1277–1287.

Moreau,V. *et al.* (2006) Discontinuous epitope prediction based on mimotope analysis. *Bioinformatics*, **22**, 1088–1095.

Motta,S. *et al.* (2005) Modelling vaccination schedules for a cancer immunoprevention vaccine. *Immunome Res.*, **1**, 5.

Moutaftsi,M. *et al.* (2006) A consensus epitope prediction approach identifies the breadth of murine T(CD8+)-cell responses to vaccinia virus. *Nat. Biotechnol.*, **24**, 817–819.

Mumey,B.M. *et al.* (2003) A new method for mapping discontinuous antibody epitopes to reveal structural features of proteins. *J. Comput. Biol.*, **10**, 555–567.

Murugan,N. and Dai,Y. (2005) Prediction of MHC class II binding peptides based on an iterative learning model. *Immunome Res.*, **1**, 6.

Nielsen,M. *et al.* (2003) Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci.*, **12**, 1007–1017.

Nielsen,M. *et al.* (2004) Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach. *Bioinformatics*, **20**, 1388–1397.

Nielsen,M. *et al.* (2005) The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage. *Immunogenetics*, **57**, 33–41.

Nielsen,M. *et al.* (2007a) Quantitative, pan-specific predictions of peptide binding to HLA- A and-B locus molecules. *PLoS ONE*, **2**, e796.

Nielsen,M. *et al.* (2007b) Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. *BMC Bioinformatics*, **8**, 238.

Noguchi,H. *et al.* (2002) Hidden Markov model-based prediction of antigenic peptides that interact with MHC class II molecules. *J. Biosci. Bioeng.*, **94**, 264–270.

Novotny,J. *et al*. (1986) Antigenic determinants in proteins coincide with surface regions accessible to large probes (antibody domains). *Proc. Natl. Acad. Sci. USA*, **83**, 226–230.

Nussbaum,A.K. *et al*. (2001) {PAProC}: a prediction algorithm for proteasomal cleavages available on the {WWW}. *Immunogenetics*, **53**, 87–94.

Odorico,M. and Pellequer,J.L. (2003) BEPITOPE: predicting the location of continuous epitopes and patterns in proteins. *J. Mol. Recognit*., **16**, 20–22.

Pamer,E.G. *et al*. (1991) Expression and deletion analysis of the Trypanosoma brucei rhodesiense cysteine protease in Escherichia coli. *Infect. Immun*., **59**, 1074–1078.

Pappalardo,F. *et al*. (2006) Analysis of vaccine's schedules using models. *Cell. Immunol*., **244**, 137–140.

Parker,J.M. *et al*. (1986) New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites. *Biochemistry*, **25**, 5425–5432.

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

Pellequer,J.L. *et al*. (1991) Predicting location of continuous epitopes in proteins from their primary structures. *Meth. Enzymol*., **203**, 176–201.

Pellequer,J.L. *et al*. (1993) Correlation between the location of antigenic sites and the prediction of turns in proteins. *Immunol. Lett*., **36**, 83–99.

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

Peters,B. *et al*. (2003) Identifying MHC class I epitopes by predicting the TAP transport efficiency of epitope precursors. *J. Immunol*., **171**, 1741–1749.

Peters,B. *et al*. (2005) The immune epitope database and analysis resource: from vision to blueprint. *PLoS Biol*., **3**, e91.

Peters,B. *et al*. (2006) A community resource benchmarking predictions of peptide binding to MHC-I molecules. *PLoS Comput. Biol*., **2**, e65.

Petrovsky,N. and Brusic,V. (2006) Bioinformatics for study of autoimmunity. *Autoimmunity*, **39**, 635–643.

Rammensee,H. *et al*. (1999) SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics*, **50**, 213–219.

Rapin,N. *et al*. (2006) Modelling the human immune system by combining bioinformatics and systems biology approaches. *J. Biol. Phys*., **32**, 335–353.

Reche,P.A. and Reinherz,E.L. (2004) Definition of MHC supertypes through clustering of MHC peptide binding repertoires. In *Artificial Immune Systems, Proceedings*. pp. 189–196.

Reche,P.A. *et al*. (2002) Prediction of MHC class I binding peptides using profile motifs. *Hum. Immunol*., **63**, 701–709.

Regenmortel,M.H.V.V. and Muller,S. (1999) *Synthetic Peptides as Antigens*. Elsevier, Amsterdam.

Revy,P. *et al*. (2001) Functional antigen-independent synapses formed between T cells and dendritic cells. *Nat. Immunol*., **2**, 925–931.

Ribeiro,R.M. *et al*. (2002) Modeling deuterated glucose labeling of T-lymphocytes. *Bull. Math. Biol*., **64**, 385–405.

Rothbard,J.B. and Taylor,W.R. (1988) A sequence pattern common to T cell epitopes. *Embo. J*., **7**, 93–100.

Rotzschke,O. *et al*. (1991) Exact prediction of a natural T cell epitope. *Eur. J. Immunol*., **21**, 2891–2894.

Saxová,P. *et al*. (2003) Predicting proteasomal cleavage sites: a comparison of available methods. *Int. Immunol*., **15**, 781–787.

Schafer,J.R. *et al*. (1998) Prediction of well-conserved {HIV}-1 ligands using a matrix-based algorithm, EpiMatrix. *Vaccine*, **16**, 1880–1884.

Schirle,M. *et al*. (2001) Combining computer algorithms with experimental approaches permits the rapid and accurate identification of T cell epitopes from defined antigens. *J. Immunol. Meth*., **257**, 1–16.

Schreiber,A. *et al*. (2005) 3D-Epitope-Explorer (3DEX): localization of conformational epitopes within three-dimensional structures of proteins. *J. Comput. Chem*., **26**, 879–887.

Seiden,P.E. and Celada,F. (1992) A model for simulating cognate recognition and response in the immune system. *J. Theor. Biol*., **158**, 329–357.

Sette,A. and Sidney,J. (1998) HLA supertypes and supermotifs: a functional perspective on HLA polymorphism. *Curr. Opin. Immunol*., **10**, 478–482.

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

Sette,A. *et al*. (1989) Prediction of major histocompatibility complex binding regions of protein antigens by sequence pattern analysis. *Proc. Natl. Acad. Sci. USA*, **86**, 3296–3300.

Smith,G.P. and Petrenko,V.A. (1997) Phage display. *Chem. Rev*., **97**, 391–410.

Sollner,J. (2006) Selection and combination of machine learning classifiers for prediction of linear B-cell epitopes on proteins. *J. Mol. Recognit*., **19**, 209–214.

Sollner,J. and Mayer,B. (2006) Machine learning approaches for prediction of linear B-cell epitopes on proteins. *J. Mol. Recognit*., **19**, 200–208.

Stoltze,L. *et al*. (2000) Two new proteases in the MHC class I processing pathway. *Nat. Immunol*., **1**, 413–418.

Stryhn,A. *et al*. (1996) Peptide binding specificity of major histocompatibility complex class I resolved into an array of apparently independent subspecificities: quantitation by peptide libraries and improved prediction of binding. *Eur. J. Immunol*., **26**, 1911–1918.

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

Sylvester-Hvid,C. *et al*. (2004) SARS CTL vaccine candidates; HLA supertype-, genome-wide scanning and biochemical validation. *Tissue Antigens*, **63**, 395–400.

Tarnovitski,N. *et al*. (2006) Mapping a neutralizing epitope on the SARS coronavirus spike protein: computational prediction based on affinity-selected peptides. *J. Mol. Biol*., **359**, 190–201.

Tenzer,S. *et al*. (2005) Modeling the MHC class I pathway by combining predictions of proteasomal cleavage, TAP transport and MHC class I binding. *Cell. Mol. Life Sci*., **62**, 1025–1037.

Tenzer,S. *et al*. (2004) Quantitative analysis of prion-protein degradation by constitutive and immuno-20S proteasomes indicates differences correlated with disease susceptibility. *J. Immunol*., **172**, 1083–1091.

Thompson,C.B. (1995) New insights into {V}({D}){J} recombination and its role in the evolution of the immune system. *Immunity*, **3**, 531–539.

Thornton,J.M. *et al*. (1986) Location of 'continuous' antigenic determinants in the protruding regions of proteins. *Embo. J*., **5**, 409–413.

Tong,J.C. *et al*. (2007) Methods and protocols for prediction of immunogenic epitopes. *Brief Bioinform*., **8**, 96–108.

van Regenmortel,M.H. and Pellequer,J.L. (1994) Predicting antigenic determinants in proteins: looking for unidimensional solutions to a three-dimensional problem? *Pept. Res*., **7**, 224–228.

van Regenmortel,M.H.V. (1996) Mapping epitope structure and activity: from one-dimensional prediction to four-dimensional description of antigenic specificity. *Methods*, **9**, 465–472.

Wang,M. *et al*. (2007) CTL epitopes for influenza A including the H5N1 bird flu; genome-, pathogen-, and HLA-wide screening. *Vaccine*, **25**, 2823–2831.

Yewdell,J.W. and Bennink,J.R. (1999) Immunodominance in major histocompatibility complex class I-restricted T lymphocyte responses. *Annu. Rev. Immunol*., **17**, 51–88.

Yu,K. *et al*. (2002) Methods for prediction of peptide binding to MHC molecules: a comparative study. *Mol. Med*., **8**, 137–148.

Zhang,G.L. *et al*. (2006) PRED(TAP): a system for prediction of peptide binding to the human transporter associated with antigen processing. *Immunome Res*., **2**, 3.

Zhang,Y. and Williams,D.B. (2006) Assembly of MHC class I molecules within the endoplasmic reticulum. *Immunol. Res*., **35**, 151–162.

Zhang,Z. and Henzel,W.J. (2004) Signal peptide prediction based on analysis of experimentally verified cleavage sites. *Protein Sci*., **13**, 2819–2824.

Zhihua,L. *et al*. (2004) Toward the quantitative prediction of T-cell epitopes: QSAR studies on peptides having affinity with the class I MHC molecular HLA-A*0201. *J. Comput. Biol*., **11**, 683–694.

Zhu,S. *et al*. (2006) Improving MHC binding peptide prediction by incorporating binding data of auxiliary MHC molecules. *Bioinformatics*, **22**, 1648–1655.