ANNUAL REVIEWS

Annual Review of Biomedical Data Science

Immunoinformatics: Predicting Peptide–MHC Binding

Morten Nielsen,[1,2] Massimo Andreatta,[2] Bjoern Peters,[3,4] and Søren Buus[5]

[1]Department of Health Technology, Technical University of Denmark, DK-2800 Kongens Lyngby, Denmark; email: morni@dtu.dk

[2]Instituto de Investigaciones Biotecnológicas, Universidad Nacional de San Martín, CP 1650 San Martin, Buenos Aires, Argentina

[3]Division of Vaccine Discovery, La Jolla Institute for Immunology, La Jolla, California 92037, USA

[4]Department of Medicine, University of California, San Diego, La Jolla, California 92093, USA

[5]Department of Immunology and Microbiology, Faculty of Health Sciences, University of Copenhagen, DK-2200 Copenhagen, Denmark

Annu. Rev. Biomed. Data Sci. 2020. 3:191–215

First published as a Review in Advance on April 27, 2020

The *Annual Review of Biomedical Data Science* is online at biodatasci.annualreviews.org

https://doi.org/10.1146/annurev-biodatasci-021920-100259

Keywords

T cells, MHC, antigen presentation, immune epitopes, machine learning

Abstract

Immunoinformatics is a discipline that applies methods of computer science to study and model the immune system. A fundamental question addressed by immunoinformatics is how to understand the rules of antigen presentation by MHC molecules to T cells, a process that is central to adaptive immune responses to infections and cancer. In the modern era of personalized medicine, the ability to model and predict which antigens can be presented by MHC is key to manipulating the immune system and designing strategies for therapeutic intervention. Since the MHC is both polygenic and extremely polymorphic, each individual possesses a personalized set of MHC molecules with different peptide-binding specificities, and collectively they present a unique individualized peptide imprint of the ongoing protein metabolism. Mapping all MHC allotypes is an enormous undertaking that cannot be achieved without a strong bioinformatics component. Computational tools for the prediction of peptide–MHC binding have thus become essential in most pipelines for T cell epitope discovery and an inescapable component of vaccine and cancer research. Here, we describe the development of several such tools, from pioneering efforts to the current state-of-the-art methods, that have allowed for accurate predictions of

peptide binding of all MHC molecules, even including those that have not yet been characterized experimentally.

INTRODUCTION

Our survival depends on an immune system that can eradicate biological threats like infections and cancer. Studying and modeling the immune system, in both health and disease, are central to the development of clinical interventions to a wide array of diseases, from cancer to infections and autoimmune disorders. As in many other fields in biology, the growing amount of immunological data generated by increasingly high-throughput methods requires the development of efficient computational methods. The emerging field of immunoinformatics sits at the intersection of immunology and computer science and aims to provide algorithms and analytical tools to aid the interpretation of immunological data and processes (1).

The ability to recognize targets of foreign origin (nonself) without attacking the tissues of the host (self) is a hallmark of the immune system and is primarily the function of T cells. T cells recognize their targets in cell–cell interactions, which involve T cell receptors (TCRs) and, as ligands, peptide–major histocompatibility complex (MHC) complexes (pMHCs) expressed by antigen-presenting cells (APCs). In the thymus, highly diverse peptide-specific, MHC-restricted TCR repertoires are generated by somatic gene rearrangements, each of which is expressed by a rare T cell clone. Through positive and negative selection, the T cells are educated to recognize self-MHC and tolerate any presented self-peptide. This creates a naïve T cell repertoire that is self-MHC restricted and prepared to recognize foreign (nonself) peptides encountered in the periphery at later times (2). Should that happen, a specific primary adaptive immune response occurs in which one or more naïve T cell clones of appropriate specificities are selected, activated, and expanded. Eventually, such expanded T cell clones contract, leaving behind persistent pools of specific memory T cells, which upon reexposure to the pathogen are rapidly recruitable and potentially afford lifelong protection.

Antigen processing and presentation involve a series of events starting with enzymatic fragmentation of the source protein antigen, selection of one or more of the resulting peptides by MHC molecules, and export of the resulting pMHCs to the APC cell membrane, where these complexes are stably displayed awaiting T cell arrival and scrutiny (**Figure 1**). $CD8^+$ cytotoxic T cells recognize peptides from the cytosol presented by MHC class I molecules, which are expressed by all nucleated cells. In contrast, $CD4^+$ T helper cells recognize peptides that have been sampled from the endocytic pathway and presented by MHC class II molecules, which are expressed by professional APCs. This dichotomy hints at the overall recognition and effector functions of the two classes of MHC molecules: Through MHC class I, $CD8^+$ cytotoxic T cells gain access to information about intracellular threats in any dividing cell and may try to eradicate infected or transformed cells; through MHC class II, $CD4^+$ T helper cells gain access to information about extracellular threats and may coordinate the response of other cells (3).

MHC classes I and II are heterodimeric, transmembrane glycoproteins forming a unique, extended peptide-binding groove that accommodates peptides (**Figure 2**). Through a few primary anchor positions, the MHC establishes a stable and broadly specific peptide interaction. A subtle but important difference exists between MHC classes I and II: Peptides bound to MHC class I tend to be short (typically 9 amino acids) and confined within the peptide-binding site of MHC class I, whereas peptides bound to MHC class II tend to be longer (typically 15 amino acids) and extend out of the peptide-binding site of MHC class II (5).

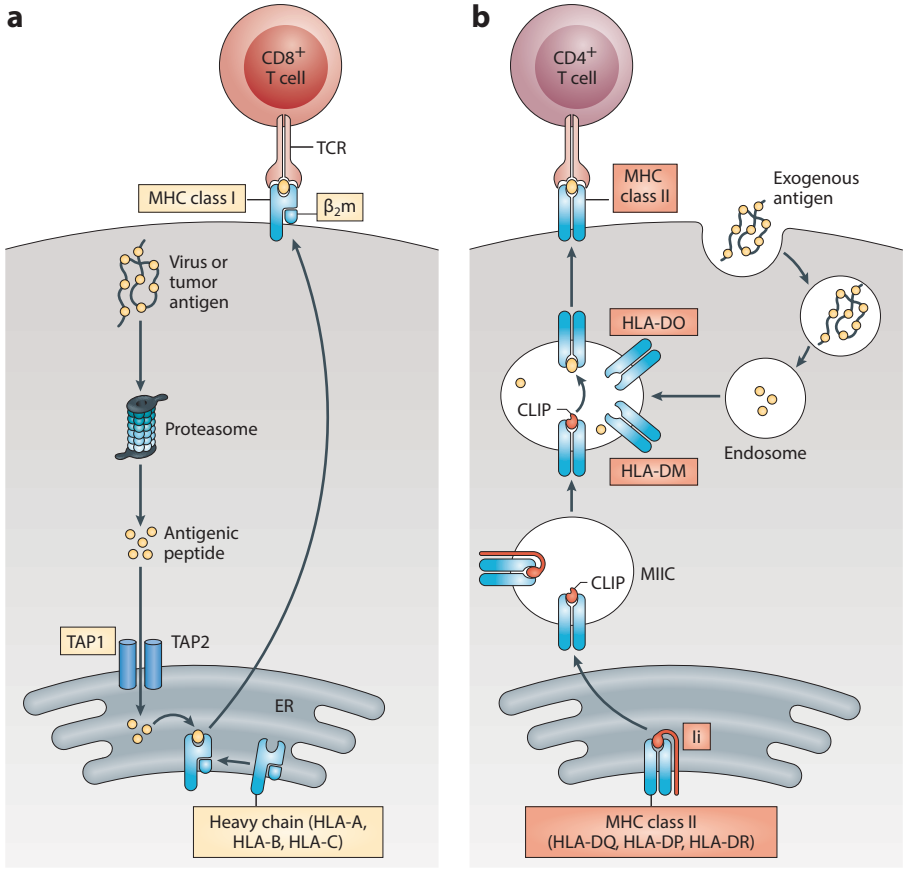

Figure 1

(*a*) The MHC class I antigen-presenting pathway. Proteins transcribed and translated are proteolytically processed in the cytosol by the proteasome; peptides are translocated into the ER via TAP and potentially undergo further trimming by ERAP prior to binding to de novo synthesized MHC I molecules in the ER; a quality control step is completed (not shown); and the resulting pMHCs are translocated to the cell surface for $CD8^+$ T cell scrutiny. (*b*) The MHC class II antigen-presenting pathway. Exogenous antigens are taken up by APCs by endocytosis, and peptides are generated by enzymatic fragmentation and traffic to MIIC. De novo synthesized MHC class II molecules bind to CLIP and traffic to MIIC. Here, controlled by HLA-DM/DO, the CLIP peptide is removed and the incoming peptides are offered to MHC class II molecules. The bound peptides are exported as pMHCs to the cell surface for $CD4^+$ T cell scrutiny. Abbreviations: APC, antigen-presenting cell; β_2m, β_2 microglobulin; CLIP, MHC class II–associated invariant chain peptide; ER, endoplasmic reticulum; ERAP, ER aminopeptidase; Ii, invariant chain; HLA, human leukocyte antigen; MIIC, MHC class II compartment; pMHC, peptide–MHC complex; TAP, transporter associated with antigen processing; TCR, T cell receptor. Panel adapted with permission from Reference 4; copyright 2012 Springer Nature.

A brief historical account of the paradigm shift to the modern view of T cell recognition is warranted. Since the early 1900s, it has been known that the outcome of allogeneic transplantation is under genetic control. This eventually led to the identification of the MHC gene complex [generically *MHC*, *H2* in mice, and *HLA* (human leukocyte antigen) in humans]. The genetic organization and polymorphisms of the HLA were established through a seminal series of international histocompatibility workshops. The importance of HLA typing and matching

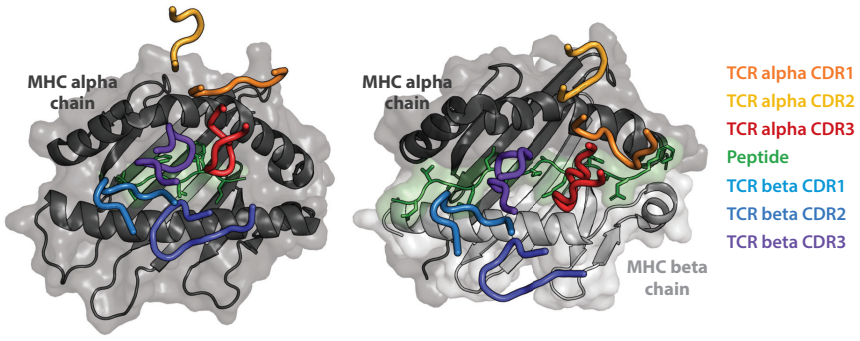

Figure 2

Structure of MHC class I and II molecules. An extended peptide-binding groove is formed between two parallel alpha helices situated over a floor made up of an antiparallel beta-sheet. The binding groove for class I is closed, whereas the binding groove for class II is open. The T cell receptor (TCR) interacts with the peptide–MHC complex predominantly via six complementarity-determining regions (CDRs), three from each chain. The two CDR3s interact mainly with the peptide. Figure courtesy of Kamilla Kjærgaard Munk.

became apparent, in particular for bone marrow transplantation, and the association between HLA and autoimmune disease was established (6). Seemingly unrelated experiments conducted in the mid-1940s hinted at antibody responses being under genetic control. Eventually, single immune-response (*Ir*) genes could be identified (7), *MHC* was found to be in close genetic linkage with *Ir* (8) [today, we know that *MHC* and *Ir* genes are one and the same (6)], and MHC was found to be in control of collaborations between B and T cells, as well as between macrophages and T cells (6). In 1974, Zinkernagel & Doherty (9) showed that virus antigens were recognized in an MHC-restricted manner. In 1978, Benacerraf (10) and Rosenthal (11) independently proposed that MHC specifically selected antigen fragments and presented them to T cells; however, the nature of MHC control of immune responsiveness remained hotly contested for almost another decade (see Reference 12 for a vivid account by Jan Klein of a period of "great, great confusion"). All of the components that are now known to control the specific interaction between T cells and APCs—the TCR, the MHC, and the antigen—were initially poorly defined, as was the cellular interaction: Did it involve one or two T cell–derived receptors recognizing one or two APC-derived ligands (antigen and MHC), either as two separate ligands, as a single complex ligand, or as an altered ligand? These puzzles were answered, one at a time: Peptides, not intact proteins, are the true antigenic ligands (13, 14); the MHC is the only other APC-derived ligand needed (15); and the TCR is responsible for both peptide specificity and MHC restriction (16).

The core question of the nature of the interaction between peptides and MHC was finally answered in a series of functional, biochemical, and structural experiments culminating around 1985–1987. In 1985, Babbitt et al. reported that appropriate peptides interacted in a specific and saturable manner with affinity-purified MHC class II molecules in a biochemical in vitro assay (17) and demonstrated that this correlated with the ability to stimulate T cells with peptide–MHC class II complexes. Shortly thereafter, Buus et al. reproduced and extended these biochemical and functional findings (18). Rather surprisingly and despite the low affinity of the interactions measured, the biochemically generated peptide–MHC class II complexes were found to be very stable. In rapid succession, it was shown that immunogenic peptides bind strongly to their MHC restriction elements, supporting the hypothesis of determinant selection (19); that different peptides restricted to the same restriction element compete for biochemical binding to that MHC (19–21) and compete for presentation to T cells restricted to that MHC (19); and that the peptide binding

site is made up of both chains of the MHC class II molecule (19). The first indications of amino acid similarities between peptides binding to the same MHC class II molecule were also provided (19–21). Observing that less than 10% of MHC class II molecules could bind the offered peptides, Buus et al. suggested that affinity-purified MHC molecules were largely preoccupied with natural self-peptides (22), an observation that was confirmed by acid elution experiments (23). Completing this paradigm shift, in 1987 Bjorkman et al. (24) crystallized an MHC class I molecule and identified a structure with a unique peptide-binding groove, which appeared to hold peptides in an extended conformation. The polymorphic amino acid positions of the MHC were predominantly found in this peptide-binding region, where they affected peptide binding and T cell recognition (25). A similar polymorphic peptide-binding structure was eventually found in MHC class II molecules (26).

Peptide binding to MHC is the single most selective event contributing to the outcome of antigen processing and presentation (27). It is currently determined using one of two complementary experimental approaches: one that investigates which synthetic peptides will bind to MHC molecules in vitro, and another that investigates what the MHC has already bound in vivo. The in vitro binding approach uses biochemical assays such as gel filtration, a robust and accurate assay (18) and, more recently, high-throughput, preferably homogenous, assays (for an overview, see Reference 28) to quantitate the binding of synthetic peptides to MHC molecules that have been purified from appropriate cell lines (28) or generated recombinantly (29). Any peptide that can be synthesized can be examined. A variant using positional scanning combinatorial peptide libraries affords a particularly comprehensive and unbiased analysis of MHC class I specificity (30). A large body of this kind of data has been deposited at the Immune Epitope Database (IEDB) (31). The in vivo binding approach uses natural peptides that have been acid-eluted off affinity-purified MHC molecules. In 1991, seminal work by Falk et al. used Edman degradation to analyze pools of eluted peptides (32). In 1992, similarly seminal work by Engelhard used tandem mass spectrometry to sequence individually eluted peptides (33).

Several earlier manuscripts and reviews have described the developments within the field of immunoinformatics and prediction of MHC antigen presentation, including References 34–36. Annually more than 30 papers are published describing novel approaches to resolve the task of predicting antigen presentation by MHC. In this review, we do not seek to provide a comprehensive overview of all the different contributions within the field, but rather provide a focused perspective of the essential discoveries and achievements, provide guidance on best practices, and outline limitations and remaining challenges. A summary of the tools described in this review is included in **Table 1**.

PREDICTION OF PEPTIDE-MHC BINDING

Simple Motif-Based Models

The observation that MHC molecules have binding preferences that could be characterized in terms of simple binding motifs led to the development of the first motif-based MHC binding prediction methods. In early work by Sette et al. (45), motifs for the two mouse MHC class II molecules, I-Ed and I-Ad, were described in terms of quantitative scores defined from position-specific propensities. Later work refined this picture of MHC binding motifs by defining anchor positions located with well-defined sequential spacing in the peptide where only a limited set of tolerated amino acids substitutions were allowed (32). These works further demonstrated how the anchor positions often were shared between different MHC molecules but that the binding motifs were unique (32). Based on this, more refined motif-based prediction schemes were defined with differential scoring of the amino acid propensity depending on the peptide position. A prominent

Table 1 MHC binding prediction methods available and described in this review

Method	URL	Pan-specific	Includes EL data	Reference
Class I				
NetMHC	**http://www.cbs.dtu.dk/services/NetMHC**	No	No	37
NetMHCpan	**http://www.cbs.dtu.dk/services/NetMHCpan**	Yes	Yes	38
MixMHCpred	**https://github.com/GfellerLab/MixMHCpred**	No	Yes	39
MHCflurry	**https://github.com/openvax/mhcflurry**	No	Yes	40
BIMAS	Decommissioned on March 8, 2019	No	No	41
SYFPEITHI	**http://www.syfpeithi.de**	No	Yes	42
SMM	**http://tools.iedb.org/mhci**	No	No	43
Class II				
NetMHCIIpan	**http://www.cbs.dtu.dk/services/NetMHCIIpan**	Yes	No	44

Abbreviation: EL, eluted ligand.

example is the SYFPEITHI prediction model (46) where binding propensity scores are estimated based on statistics from MHC class I eluted-ligand data (EL data). Using direct binding assays, Stryhn et al. employed positional scanning combinatorial peptide libraries to identify both the MHC class I anchor positions and propensity scores of the different amino acids at each peptide position, thus providing a full characterization of the MHC binding motif (30). Alternatively, the scores were based on measured binding data from single-substitution analogs of known ligands (47). Eventually, both the eluted ligand approach (48) and the single-substitution approach were extended to MHC class II binding ligands (49).

Toward the First Machine Learning Models

While these pioneering motif-based prediction methods enabled the first cataloging and classification of MHC binding preferences, it soon became clear that they very often suffered from low sensitivity and hence often failed to identify large proportions of validated MHC-binding peptides. Inspired by advances in machine learning, the mid- to late 1990s saw the development of a second generation of prediction models. In this kind of approach, models were trained on experimental data, aiming to minimize the error between the predictions of the model and a set of experimental measurements. The resulting mathematical model could then be applied to make predictions on new data: in this case, new peptide–MHC pairs of interest.

The BIMAS model proposed by Parker et al. was the first example of a predictor trained directly on experimental data, fitting a matrix model using linear regression (LR) on the measured half-life of HLA-A2 complexes with bound peptide (41). Later, more complex models were proposed, including artificial neural networks (ANNs) (50–55), hidden Markov models (HMMs) (56, 57), and QSAR (quantitative structure-affinity relationship)-based regression models (58).

A common factor limiting the success of these models was the insufficient availability of experimental data characterizing peptide binding to the investigated MHC molecules. By way of example, the ANN model proposed by Milik et al. (51) was trained on merely ∼200 data points. Because machine learning models built on small datasets often contain more parameters than can be confidently estimated from the data, the performance of the early machine learning methods often suffered from a phenomenon generally known as overfitting. Nonetheless, some of the simpler models such as BIMAS have, over the years, demonstrated a high performance in epitope prediction.

Larger Peptide–MHC Databases: More Reliable Prediction Methods

With the formulation of novel, high-throughput peptide–MHC-binding assays (see above) and with the creation of large publicly available databases covering both a broader set of MHC molecules and a more in-depth characterization of individual MHC molecules [SYFPEITHI (46), MHCBN (59), MHCPEP (60), and IEDB (31)], it became possible to develop and validate more complex prediction models, including methods based on machine learning. Early examples of such models included LR models such as stabilized matrix method (SMM) (43, 61) and higher-order regression (HR) models such as SVMHC (62), SVRMHC (63), and NetMHC (55). Here, the HR models allow incorporation of correlations between neighboring residues and are hence expected to better capture nonlinear relationships between peptide sequence and MHC binding.

With the growing list of available methods, the lack of objective benchmarks and criteria for the evaluation of predictive performance became critical. Peters et al. (64) published one of the first attempts to rigorously assess the state of the art in the field by comparing the performance of a panel of previously published and publicly available methods to the performance of three methods, ARB (average relative binding), SMM, and NetMHC, that had been retrained and benchmarked on common data. One important conclusion from this benchmark study was that SMM and NetMHC (called ANN in the benchmark) overall outperformed all other methods included in the analysis. Furthermore, the retrained versions of the tools outperformed the earlier versions of the same methods, strongly suggesting that continuous retraining of prediction methods as additional data become available is essential to maintain optimal performance. Another observation was that the SMM and NetMHC methods achieved comparable performance for larger datasets, suggesting that the abovementioned nonlinear correlations play a minor role in defining the peptide–MHC interaction. Peptides bind to the MHC in an extended conformation (**Figure 2**) (65), and as a first approximation, the strength of the interaction should depend on the sum of the binding energies of the individual amino acids, suggesting that peptide–MHC binding is mostly a linear problem. This may explain why simple LR methods have achieved high predictive performance for the peptide–MHC system, outperforming more complex structure-based methods (66). Later benchmark studies demonstrated that neural network–based methods, such as NetMHC, can indeed improve the prediction of T cell epitopes (67). However, these neural network methods were all shallow (most often limited to a single hidden layer), further supporting the notion that nonlinear contributions to peptide–MHC binding are limited.

In parallel with the development of accurate prediction methods for the MHC class I system, similar efforts were dedicated to the prediction of peptide binding to MHC class II molecules. The open binding groove of MHC class II, which allows longer peptides to protrude out of the binding cleft (68), makes peptide binding predictions of MHC class II considerably more challenging compared to MHC class I. While peptides interact with MHC class II using a binding core of nine residues, they are generally out of frame in terms of the location of this binding core. Therefore, to ensure accurate model training, peptides must be aligned to a common binding core. The list of machine learning frameworks proposed to solve this challenge is long, including HMMs (69), SVMs (support vector machines) (70, 71), Gibbs sampling (72), and ANNs (53, 73), among others. An important algorithmic development toward accurate modeling of pMHC II binding was the NNAlign method (**Figure 3**) (73, 74). Using an iterative loop over a set of training peptide data with measured MHC binding values, NNAlign identifies the optimal binding core of nine amino acids and at the same time predicts binding for a given peptide given the current model parameters; next, it updates the model parameters to minimize the difference between the predicted and measured binding. Through this iterative procedure, the algorithm achieves simultaneous alignment and binding motif characterization.

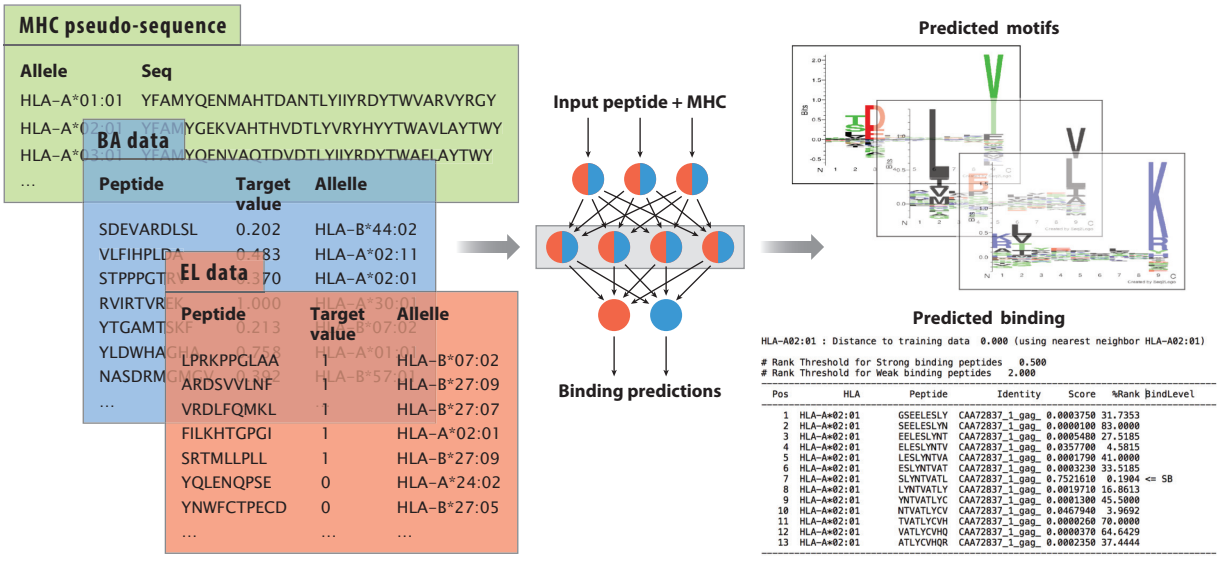

Figure 3

The NNAlign machine learning framework. Different kinds of peptide–MHC binding data are integrated in a machine learning framework leveraging information between multiple MHC molecules and peptide length, resulting in a pan-length, pan-MHC prediction method that captures individual binding motifs and allows for accurate epitope prediction. Abbreviations: BA, binding affinity; EL, eluted ligand.

Benchmark studies soon demonstrated that, in addition to MHC class I, ANN-based models also achieved the highest predictive power for class II (75). Over the years, the NNAlign framework (73) has been refined continuously, and it now serves as the main framework for training NetMHC and NetMHCII (as well as NetMHCpan/NetMHCIIpan, further discussed below) (44, 76, 77).

An important shortcoming of the early machine learning prediction methods for MHC class I peptide binding was that they (as most other early machine learning methods) required the input data to be of uniform length. Consequently, individual prediction methods had to be developed for each peptide length separately. This severely limited the predictive power of these tools since very limited binding data were available for peptides with lengths outside the canonical range of 9–10 amino acids. As a workaround, approximation methods were used to assess the binding of peptides of a length that was different from that used to train a given prediction model. For instance, Lundegaard et al. (78) proposed a simple approximation where prediction models trained on 9-mer peptides were used to predict binding of peptides of length 8, 10, and 11. This approximation assumes that peptides of length other than 9 are accommodated through a structural change maintaining the MHC anchor binding preferences while adapting to the closed MHC class I binding pocket. For longer peptides, this results in a binding mode where one or more (depending on the peptide length) consecutive residues bulge out of the binding cleft, whereas 8-mers assume a more extended conformation. By representing these different binding modes as deletions (in the case of bulging) and insertions (in the case of extension), one can calculate the affinity of any peptide of length 8–11 as the average predicted affinity of all possible pseudo-9-mer binding modes.

Whether length-specific modeling or approximation methods were used, the results were a lower accuracy for predictions of non-9-mer peptides (37). Moreover, these approaches could

account for neither the length distribution of MHC-bound peptides nor the peptide length preference variations between individual MHC alleles. For example, in the earlier versions of NetMHC, predictions for non-9-mer peptides were extrapolated from prediction models trained on 9-mer peptides using the approximation as described above. As a result, on average, the method predicted the same number of binders for each peptide of length 8–11. This was in clear contrast to the length preference observed for epitopes and naturally presented HLA binders (79, 80), where the vast majority of peptides are 9 amino acids long, leading to a large proportion of false positives among the predicted non-9-mer binders.

Aiming to address the differential peptide length preferences of MHC class I molecules, Trolle et al. (80) suggested an empirical length-dependent correction of the peptide prediction score. This length correction effectively adjusted the modeled peptide length preference to that observed in naturally presented HLA binders and greatly improved the identification of T cell epitopes. Concurrently, advances in the NNAlign framework allowed it to train prediction methods based on datasets of different peptide lengths (37). In this case, however, the reconciliation of peptides of any length onto the 9-mer binding mode was performed during training, allowing peptides of all lengths to be included in the construction of the model and effectively enabling the development of a pan-length prediction method. In addition, the method allowed for binding modes with C- or N-terminal extensions—a phenomenon suggested to occur with a small, but non-negligible, frequency for a list of HLA-A and HLA-B molecules (examples reported in References 81–83). This extension to the NNAlign framework afforded two important advantages. Firstly, training single models on datasets covering different peptide lengths allowed the method to leverage information obtained from datasets representing peptides of different length. Secondly, but perhaps more importantly, the pan-length approach allowed the peptide length profile to be learned for different MHC molecules. Later, an alternative approach for pan-length model development was proposed by O'Donnell et al. (40). In their MHCflurry model, peptides of variable length were transformed using a fixed-length 15-mer peptide and an encoding designed to preserve the location of MHC anchor positions. Several benchmark studies have demonstrated comparable performance between these two approaches (84–87).

Selecting a Threshold to Define Binders

Another critical issue faced when applying (predicted or measured) MHC binding as a filter for rational epitope discovery is the definition of the binding classification threshold to be used to identify immunogenic T cell epitopes. Early work by Sette et al. (88) suggested that a measured binding threshold of 500 nM captured most of the immunogenic $CD8^+$ T cell epitopes for HLA-A*02:01. Likewise, Southwood et al. (89) analyzed the relationship between measured binding affinity and T cell immunogenicity for a panel of HLA-DR molecules and found that a measured binding threshold of 1,000 nM captured most of the immunogenic $CD4^+$ T cell epitopes. However, it became clear that such a universal binding threshold was suboptimal for epitope discovery, as several studies could demonstrate that different HLA molecules shared peptide binding repertoires of varying sizes (90, 91) and presented peptides in complex with MHC molecules at very different affinity values (92). This effect was demonstrated when assessing MHC binding using both experimental binding assays and prediction models (90, 92). To reconcile different affinity thresholds between different HLA molecules, we have suggested that, for immunogenicity classification purposes, percentile rank scores rather than binding affinity values should be used as thresholds (93). In short, percentile rank scores are calculated as the percentage of peptides (in a precalculated score distribution derived from a large set of random natural peptides) that have a better score than the peptide in question. For example, a percentile rank score of 0.1% indicates

that only one out of a thousand random peptides is expected to obtain a prediction score better than the query peptide. In practice, rank scores act as a normalization of binding affinity scores over the underlying score distribution of the individual MHC molecule and are therefore useful when equating and comparing different MHCs. Subsequent work has demonstrated how the use of such rank scores, rather than predicted binding affinity values, led to an overall higher predictive power when analyzing the peptidome repertoires in the context of multiple HLA molecules (94). In this context, Paul et al. (92) went further and demonstrated, for a limited set of HLA molecules, that using an experimentally defined allotype-specific binding threshold was optimal. The debate on this issue is still open, but the consensus of the field is leaning toward using percentile rank score for epitope classification.

Dealing with HLA Polymorphism

All the prediction methods discussed up to this point have been allotype specific, meaning that a separate prediction model needs to be developed for each individual MHC molecule. Since 11,405 HLA I allotypes are described in the current version 3.37 of the IMGT (International Immunogenetics Project)/HLA database (95), it would be a very large undertaking to generate binding data characterizing each of them. The extreme polymorphism of the MHC is a major constraint in the development of allotype-specific prediction methods covering all human MHC molecules, a goal that otherwise would be of great interest for rational epitope discovery and personalized vaccine/immunotherapy design.

A first approach to solve this problem was based on the realizations that MHC molecules can be clustered into groups (so-called HLA supertypes) of molecules that bind largely overlapping peptide repertoires (96, 97) and that population-wide epitope discovery hence could be achieved by predicting binding to individual MHC molecules representing each supertype. While this approach was somewhat successful, it also became clear that the supertype concept was an oversimplification. Although MHCs within a given supertype share overlapping peptide repertoires, the overlap is far from perfect, and the success of epitope discovery could be greatly improved by using prediction methods matching the exact MHC allotype of the population/patient of interest rather than a supertype representative (93). A crucial advance toward achieving full and accurate coverage of the MHC space was the development of NetMHCpan, the first pan-specific method for MHC class I (98). NetMHCpan was inspired by the work of Brusic et al. (99), who complemented the peptide binding information used to train the prediction model with information about the amino acids defining the MHC binding cleft. This made it possible to leverage information between MHC molecules and, for the first time, to generate accurate predictions for allotypes with limited or even no binding data. Other pan-specific approaches have subsequently been proposed, such as ADT (adaptive double threading; 100), KISS (kernel-based inter-allele peptide binding prediction system; 101), and PickPocket (102), each implementing different representations of the MHC binding environment to allow for the development of pan-specific prediction models. For class II, the first pan-like prediction model was TEPITOPE (103), which enabled peptide binding prediction for a library of 51 HLA-DR molecules. TEPITOPE achieved this by using virtual matrices based on similarity between binding pocket residues from a small database of experimentally determined binding pocket profiles. Other prediction models for class II, based on approaches similar to that of NetMHCpan described above, include MultiRTA (104), MHCII-Multi (105), and NetMHCIIpan (44, 76, 77, 106), the last of which covers all class II proteins of known sequence, thus enabling true pan-specificity. Independent benchmarking has subsequently demonstrated the superior performance of the NetMHCpan/NetMHCIIpan methods for prediction of peptide binding, MHC ligands, and T cell epitopes (107–109).

The Role of MHC Binding Stability

The methods described so far have primarily been constructed to predict the peptide binding affinity of the MHC. However, MHC molecules must not only bind the peptides generated inside the cell but also retain them at the cell surface long enough to be available to rare circulating T cells of the appropriate specificity. It has been argued that the stability of the peptide–MHC interaction, rather than binding affinity, is a more relevant property to predict T cell immunogenicity (110, 111). The main reason why pMHC I stability has not been used more extensively is related to the cumbersome or low-throughput nature of current biochemical methods used to measure the dissociation of pMHC I complexes (111). In 2011, Harndahl et al. proposed an assay to resolve this experimental bottleneck (112). Inspired by the earlier work of Parker et al. (41), they radiolabeled β_2 microglobulin (β_2m) rather than the peptide and used β_2m dissociation as measured by a scintillation proximity assay to accurately monitor peptide dissociation. Using this assay, they could demonstrate that peptide–MHC stability rather than peptide affinity is a better predictor of CTL immunogenicity (111). These data were later used to construct predictors of peptide–MHC binding stability, and benchmarking of these predictors further suggested that binding stability plays an important role in defining peptide immunogenicity (113, 114). At present, datasets of peptide-MHC stability remain limited in terms of amount of data, peptide diversity, and MHC alleles covered. This, combined with a shifting focus toward mass spectrometry (MS) for sequencing immunopeptidomes (see below), has halted further the development and a wider application of prediction methods based on peptide–MHC stability in T cell epitope discovery.

Methods Predicting MHC Antigen Presentation

Peptide binding to MHC is arguably the single most selective event in antigen processing and presentation. Nonetheless, other events can affect the availability of peptides for T cell recognition. Prediction methods have been developed for proteasomal cleavage (115, 116) and TAP (transporter associated with antigen processing) efficiency (117, 118). Two important observations can be made from these studies. Firstly, the steps involved in antigen processing have specificities that can be learned and applied to identify MHC ligands. Secondly, the events of proteasomal cleavage and TAP transport are much less specific compared to that of MHC binding. Different methods integrating the prediction of the various steps in antigen processing and presentation have been proposed, including NetCTL (119), NetCTLpan (93), and MHC-pathway (120) (reviewed in Reference 107). However, only minor improvements in predictive power have been obtained (93). A comparison of the specificity of the proteasome, TAP, and MHC suggested that the MHC molecules appear to have (co)evolved to accept binding of peptides generated by the proteasome and translocated by the TAP molecules (115). Taken together, these results suggested that, rather than serving as a specificity filter, the major concerted role for the proteasome and TAP in the context of MHC class I antigen presentation is to deliver peptides of the appropriate length for MHC binding. That is, the proteasome predominantly generates short peptide fragments with an average length of 7–8 amino acids (121, 122), TAP preferably translocates peptides of length 11 (118), and the combination of these two length preferences as first approximation results in a length distribution centered around 9–10 amino acids—in agreement with the observed peptide length distribution of presented MHC ligands (79, 80). Note that the specificity of other proteases such as ERAP1 and ERAP2 can also influence the repertoire of ligands available for MHC binding (123, 124), but they have not been included in any of these prediction approaches.

Mass Spectrometry Characterization of MHC Ligandomes and Prediction Methods Trained on Mass Spectrometry Ligandome Data

The peptide repertoire presented by MHC molecules on the cell surface is commonly referred to as the MHC ligandome or immunopeptidome. In the past decade, technological advances in proteomics and MS have enabled the study of such MHC ligandomes (EL data) at an unprecedented scale and level of detail and have established a new way of studying peptide–MHC presentation (reviewed in Reference 125). As cells normally express multiple types of MHC molecules, a key challenge for the interpretation of immunopeptidome data is to identify the different MHC specificities and assign/annotate the individual peptides to one or more of these specificities. Several experimental approaches have been proposed to solve this task, including the use of monoallelic cell lines (126) and cell lines expressing a secreted form of specific MHC molecules (127). However, these approaches might not always be feasible, and it would obviously be more desirable to analyze and interpret EL data obtained from cell lines, and eventually patient samples, that express several MHC allotypes. In this context, a pioneering method, GibbsCluster, was proposed by Andreatta et al. (128). As input, GibbsCluster takes a list of peptide sequences (potentially of variable length) and uses a heuristic search to group them into clusters by optimizing the peptide similarity within clusters and the dissimilarity between clusters. Besides the sequence motif defining each cluster, additional properties such as the ligand length distribution of each cluster can be analyzed. Bassani-Sternberg et al. (129) demonstrated how this method could effectively be used to deconvolute and characterize MHC class I ligand data. Later, Bassani-Sternberg & Gfeller (39) proposed a similar method, MixMHCp, with comparable performance and predictive power. Unfortunately, and irrespective of which method is applied, it is not always possible to deconvolute the complete number of MHC specificities expressed in a given cell line, especially if some of the different MHC molecules have overlapping binding motifs or have highly varying expression levels. Moreover, assigning specific MHC molecules to the different clustered solutions relies on manual annotation and prior knowledge of allotype-specific MHC binding (130, 131). Gfeller et al. suggested an elegant, unsupervised solution to this problem by deconvoluting and automatically annotating HLA I motifs based on co-occurrence of alleles across large MHC ligand datasets (132), thereby allowing the MHC ligands to be associated with their putative MHC restriction elements.

Whichever of the above experimental or computational solutions is adopted, the outcome is an EL dataset with putatively annotated MHC restriction. Such data are a rich source of information for learning the rules of MHC-mediated antigen processing and presentation and for identifying potential T cell epitopes for both class I and class II. The amount of EL data available in the public domain is large (and rapidly growing); however, a very large body of complementary information exists in the form of conventional peptide binding data. By way of example, as of August 8, 2019, IEDB contains close to 1,200,000 MHC ligand data points. Of these, more than 300,000 are derived from MHC binding assays, and the remaining are from MHC ligand elution assays. Investigating the subset of these data characterized by single, high-resolution-typed MHC restriction, and limiting the binding affinity data to quantitative assays, the two datasets individually cover around 140 distinct MHC molecules. However, only 80 of these are shared between the binding affinity (BA) and EL data types. Given this complementarity, it is attractive to develop prediction algorithms that can benefit from both data types. One such approach was proposed by Jurtz et al. (38), who used a novel neural network architecture that integrated BA and EL data into a single training scheme, allowing information to be leveraged across the two data types. This resulted in a machine learning method that could learn (and predict) both the binding affinity of a given peptide and its likelihood of being an MHC ligand. This modeling framework, which resulted in the NetMHCpan-4.0 method, was shown to achieve predictive performance beyond that of

models trained on each data type (BA or EL) separately for both for class I and class II (38, 133). An alternative approach of integrating BA and EL data was later implemented by the MHCflurry tool (40), where the difference between qualitative EL and quantitative BA data was handled in the data presentation to the machine learning method. The results of this approach further supported that combining BA and EL data generated superior prediction models and in particular demonstrated improved performance for prediction of EL data.

As stated above, a key challenge and limiting factor for the interpretation and use of EL data is the deconvolution step required to assign each ligand to its putative MHC restriction element(s). Recently, we have proposed a framework that allows complete MHC peptidome deconvolution of EL data and, simultaneously, its automatic annotation to individual MHC molecules (85). The framework is inspired by the work of Gfeller et al. (132) and is an extension of the NNAlign neural network framework described above. The method, termed NNAlign_MA, is capable of taking single-allele datasets (peptides assigned to single MHCs) and multiallele datasets (peptides with multiple options for MHC assignments) as input and fully deconvoluting the individual MHC restriction of all peptides while simultaneously training a pan-specific MHC binding predictor. Benchmark studies of the method have demonstrated an improved performance of this framework compared to other state-of-the-art methods for prediction of MHC eluted ligands and T cell epitopes for both MHC class I and II. Importantly, this method could effectively expand the knowledge base of MHC molecules with characterized binding motifs (85).

EL data inherently contain information about the antigen processing preceding MHC presentation. For MHC class II, analyses of large-scale EL datasets have revealed the presence of clear C- and N-terminal motifs consistent with specific proteolytic cleavage signals (131, 133), and incorporation of these cleavage signals has been demonstrated to boost prediction of MHC class II ligands (133).

Exceptions to the Canonical Rules of MHC Antigen Presentation

The availability of large and diverse ligand datasets covering a broad range of different MHC molecules from a diverse list of species, including humans, nonhuman primates, mice, cattle, and swine, has allowed the field to investigate in great detail the rules that define antigen presentation, especially for MHC class I. The striking conclusion from these analyses is that the rules are surprisingly simple and largely consistent across mammalian species. In the vast majority of cases, the selection for antigen presentation is predominantly governed by the MHC [in nonmammalian species like chicken, this selection is also affected by polymorphic genes such as *TAP* and tapasin (*TAPBP*) controlling antigen processing and peptide loading (for a review, see Reference 134)], and the rules for binding to MHC across species are defined by MHC-specific binding motifs characterized by anchor positions with a well-defined spatial separation shared among the majority of MHCs (P2 and PΩ for class I, and P1, P4, P6, and P9 for class II). Historically, there has been some controversy about whether all presented peptides adhere to these rules. However, recent analyses of extensive MHC ligand datasets, and reanalysis of historic epitope data, have confirmed that the vast majority of the outliers contain nested peptides that do adhere to the rules (82, 135). This being said, analyses have also revealed a minor MHC ligand population that follows noncanonical modes of MHC antigen presentation. These noncanonical modes include presentation of ligands extending beyond the C terminus (and to a lesser degree N terminus) of the class I binding cleft (81–83), ligands with 8- or 10-mer cores for binding to MHC class II (136), and nongenomically templated peptides generated by proteasomal splicing of protein fragments (137, 138). While the true nature of these exceptions to the conventional rules of MHC antigen presentation has been indicated in several studies, the vast majority of currently characterized T cell epitopes adhere

to the canonical rules. In terms of MHC presentation, it is important to underline that current peptide–MHC binding prediction models such as NetMHC, NetMHCpan, and NetMHCIIpan all provide the possibility of incorporating and learning from peptides with noncanonical binding modes as more data supporting their biological relevance become available.

BENCHMARKING

As in most of computational biology, evaluating different MHC binding prediction algorithms relies on the availability of metrics and datasets that allow researchers to benchmark and compare their performances (139). Such benchmarks are valuable for both users, who are getting guidance on what algorithm to apply, and developers, who can objectively demonstrate if they have found a superior approach. Ideally, such benchmarks should be conducted on new data, meaning that the benchmark data were not used in the development and training of the evaluated algorithms. Assembling large enough experimental datasets for this purpose can be costly and time consuming, and keeping such data from public view just for benchmarking purposes is not justifiable. Automated benchmarks for MHC class I and MHC class II binding predictions have been added to the IEDB website to circumvent this problem; these benchmarks run all data that are newly included in the weekly database releases through a set of prediction tools and compare the predicted and measured binding assessments (108, 109). Any tool developer can register to include their tool in these benchmarks. While each of these microbenchmarks tends to have few data points, aggregating their results over time should reflect the overall reliability of different algorithms. One issue that has arisen with this automated approach is that new experimental datasets that characterize peptide–MHC binding in an appropriate way for inclusion in these benchmarks are published much less frequently than one would hope for. Many datasets only include positive data points or do not report any quantitative measurements at all. Furthermore, as not all prediction methods can make predictions for all MHC alleles or all peptide lengths, comparisons between different methods are nontrivial. While this makes the results of these automated benchmarks less definitive than one would hope for, the general principle of automating the evaluation of prediction algorithms on newly available data is an important addition to the traditional approach of having tool developers demonstrate their algorithm performance as part of their tool publication.

T CELL EPITOPE DISCOVERY

While prediction of peptide–MHC binding is an interesting problem in itself, the ultimate goal of most real-life applications is the prediction and identification of T cell epitopes. Most pipelines for T cell epitope discovery include prediction of either MHC binding or likelihood of MHC antigen presentation (representative examples include References 140–143). Given this, it is important to assess the performance of different prediction methods in terms of predicting T cell epitopes and to define the optimal modus operandi for each method. While substantial efforts have been dedicated to assessing the power of current tools' ability to predict MHC–peptide binding (see above), limited work has been published evaluating available methods for their predictive power relative to T cell epitope discovery. However, the current consensus in the field is that the following methods are leading: MixMHCpred (39), MHCflurry (40), and NetMHCpan-4.0 (38) for MHC class I, and NetMHCIIpan (44) for MHC class II. For class I, these methods have been applied and benchmarked in a series of recent publications (36, 84–87), each demonstrating a very high and comparable performance for T cell epitope identification. As discussed above, the recommended use of these methods is to select predicted epitopes by use of a percentile rank threshold. The value of this threshold depends on the scope of the given application. If the goal is to find some epitopes

(not necessarily all) and avoid spending excessive resources on false positives, a stringent threshold of 0.25% or 0.5% rank should be used. Benchmark studies including References 38, 86, and 144 have demonstrated that this threshold will identify ∼70% of the epitopes while discarding up to 99.5% of nonimmunogenic peptides. In contrast, if one is interested in identifying all epitopes—or in avoiding any false negatives (i.e., for avoidance of immunogenic biosimilars)—a less stringent threshold of 2% rank should be selected. Using such a threshold will ensure that the vast majority of epitopes are identified (∼95%), albeit at the expense of a loss in specificity (38, 86).

In the context of immunogenicity, where there is a need to identify a few epitopes out of many thousands of peptide candidates, even a specificity of 99% translates into a relatively high proportion of false-positive predictions, and the false-discovery rate (FDR) in epitope discovery projects rarely falls below 50%. This has contributed to the notion that epitope prediction remains a daunting problem (145). However, these FDRs should be appreciated in light of the fact that peptide immunogenicity is in general a rare event. Less than 0.05% of randomly selected peptides within a given pathogen are immunogenic (27); this would result in an FDR of 99.95% if peptides were selected at random. Seen in this light, an FDR of, for instance, 75% is a large improvement. However, from an applied perspective, this might still be prohibitive. Several issues are at play in explaining the high proportion of false predictions. Yewdell & Bennink (27) suggested that the TCR repertoire only matched 50% of the presented pMHCs; thus, even a perfect predictor of presented pMHCs would, in terms of immunogenicity, never achieve an FDR lower than 50%. When focusing on single proteins with known epitopes, current peptide–MHC binding prediction tools identify the majority of ligands and epitopes within the top 0.5% of the peptides of the source proteins (38, 86). However, studies suggest that not all proteins are available for MHC antigen presentation; if not accounted for, this could in itself lead to a high FDR (131). It is not fully understood what dictates this differentiation between the proteins that provide peptides for antigen presentation and those that do not, but besides protein expression and abundance (146, 147), properties related to protein degradation and translation efficiency are likely involved (131). MS has been suggested as an approach to limit the FDR, either as a means to identify proteins available for MHC antigen presentation or as a strategy for antigen discovery (148, 149). While such approaches are powerful, it is critical to realize that the gain in specificity provided by MS comes at a very high price in terms of loss of sensitivity. Even with the most recent technological (150) and computational (151, 152) advances, MS studies of peptides eluted off MHC only capture a small proportion of the set of peptides presented by MHC, and most often fail to identify a substantial proportion of validated epitopes (148, 149, 153). Given this, we believe that any rational, real-time T cell epitope discovery will remain dependent on in silico prediction methods (148).

Studies have suggested that properties of the peptide side chains facing out from the MHC binding groove can favor interaction with the TCR, and such properties have been implemented into T cell epitope prediction models with some success (154, 155). Likewise, studies have investigated how the similarity between MHC-presented peptides and the self-peptidome (144, 156) and microbiome can impact the likelihood of a relevant TCR being present in the T cell repertoire of a given individual due to positive and negative T cell selection (157). The common conclusion of these studies is that, while some deselection of peptides with high self-similarity could be observed, the predictive power of the proposed similarity models has been modest at best. This suggests that we still have a very inadequate picture of the rules of TCR cross-reactivity that define functional similarities between peptides.

Another critical issue faced when doing rational epitope discovery is the inherent diversity of the pathogens and the resulting host immune responses; rarely will different members of a population share the same HLA allotypes and be infected with the same pathogen strains. Different approaches have been suggested to deal with this problem, including focusing on binding to sets

of prevalent and functionally different HLA alleles [using for instance the HLA supertypes (96, 97)] and selecting peptides from genomic regions conserved across pathogenic strains (158). Other approaches such as Mosaic, OptiTope, Episelect, and PopCover (reviewed and benchmarked in Reference 159) have explicitly dealt with the pathogen or HLA diversity by selecting peptides that in concert provide a broad HLA and pathogen strain coverage.

DISCUSSION AND CURRENT CHALLENGES

In 1999, we proposed a Human MHC Project describing and predicting peptide binding for all human MHC molecules (160). Covering even the 413 classical HLA class I molecules registered in the contemporary version 1.1 of the IMGT-HLA database was a challenging and distant goal. The past decades have seen significant improvements in how peptide–MHC binding data are addressed experimentally. However, the number of registered molecules has increased to 11,405 (version 3.37, July 2019) and is still growing, and the development of pan-specific predictors is the only reason why the peptide binding characteristics can be said to have been solved for all human MHC molecules, now and in the future.

In this review, we have outlined the development of these highly successful prediction methods. Not only do they cover all current and future MHC molecules, but they have also achieved levels of accuracy that are second to none compared to computational predictors in other areas of biology. Nonetheless, there is still a lot to do: Current predictors have important limitations, including significant aspects of antigen presentation or recognition that are covered poorly, or not at all. There are undoubtedly important new biological discoveries waiting to be made and included in future predictors. Some examples are given below.

The predictive power of a machine learning tool is only a good as the data used for training. Posttranslational modifications (PTMs) such as phosphorylation, glycosylation, deamidation, etc. are known to influence the specificity of MHC binding and presentation (161–163); however, very limited data and no reliable prediction methods incorporating PTMs are currently available. Similarly, there is a known underrepresentation of the amino acids cysteine in MHC ligand datasets (132), leading to low predictive power for cysteine-containing peptides. Different corrective measures have been suggested to correct for this, including replacing cysteine with an X when representing a peptide sequence in NetMHC predictions (144).

All models described in this review are very simple in terms of the underlying machine learning framework: All are either simple linear-matrix-based (MixMHCpred) or feedforward neural network (the NetMHC suite and MHCflurry) models. Over the last few years, the field of machine learning has been revolutionized by the development and application of so-called deep neural network methods combining different network functions such as CNN (convolutional neural network) layers, LSTM (long short-term memory) layers, and FFNN (feedforward neural network) layers into deep (i.e., containing multiple layers) and complex network architectures. While such deep methods have proven useful within bioinformatics and biology in general (164, 165), their impact on the prediction of peptide–MHC binding has been limited (36, 86, 166). The meager benefit may be attributed to the binding of short peptides in an extended conformation, as described earlier, which largely can be approximated as a linear system, as attested by the satisfactory performance of shallow neural networks and even matrix-based prediction methods. However, future work and further independent benchmarking are needed to fully evaluate the potential of deep methods in the context of predicting MHC antigen presentation and T cell epitopes.

Several aspects of the underlying biology underlying MHC antigen presentation are poorly understood and described. One example is the abundance of protein antigens available for processing and presentation (146, 147). Studies of MHC EL datasets have suggested that high protein

expression [as measured by RNA sequencing (RNA-seq) expression] can compensate and allow for antigen presentation of relatively weak MHC-binding peptides (126, 167). However, how to optimally integrate the information contained within RNA-seq data is not trivial, and no publicly available prediction method currently offers the integration of expression data. Another example is the recent discovery of the TAPBPR quality control mechanism, which is believed to ensure that only peptides stably bound to MHC class I are exported to the APC surface and presented (168).

T cell recognition requires that pMHCs are matched by an appropriate TCR in the repertoire of the individual in question. This repertoire is uniquely shaped in a self-referential manner through thymic events that are controlled by MHC and by the specificity and cross-reactivity of the TCR, the latter of which we know very little. Finely characterizing the rules of T cell recognition may be the single most important new development to be done. Ideally, one should aim to generate recombinant (preferably soluble) TCRs, establish efficient trimolecular TCR:peptide:MHC binding assays, generate large columns of such interaction data, and construct the corresponding predictors. If this could be done, then the entire T cell repertoire of an individual could be modeled. This goal is already being pursued. Analyses of TCR data suggest that TCRs sharing common cognate peptide–MHC ligands share tractable common sequences (169, 170) and structural properties (171), in principle allowing for the prediction of T cell ligands (172, 173). Currently, the performance of such prediction models is severely hampered by the limited amount of data available. The majority of the data are, at present, limited to the CDR3 segment of the TCR beta chain. Lanzarotti et al. (171) have demonstrated that this gives an inadequate view of TCR specificity, suggesting that paired TCR alpha and beta sequence information will be needed. Recent technological advances have greatly improved researchers' ability to obtain such paired sequence information from epitope-specific T cells (174). As more of this kind of data become available, it is very likely that reliable TCR:peptide:MHC binding predictions will become a reality.

Another T cell phenomenon, immunodominance, is also poorly understood. This is a situation where the immune response in a given patient is predominantly focused toward a limited subset of the available T cell epitopes, causing a hierarchy of immune responses in terms of magnitude and prevalence. It is not clear what defines this hierarchy. Differential features of HLA antigen processing and presentation, and properties of the T cell repertoire shaped by infection history, precursor frequencies, and the VDJ (variable–diversity–joining) germ lines of the individual, are likely to play a critical role (175, 176).

Addressing these unsolved questions will require continued basic research on antigen processing and presentation, as well as on T cell recognition and responses. This will go hand in hand with technological developments supporting large-scale and high-throughput assays of the same.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This work was supported through funding from NIH (National Institutes of Health) contract 75N93019C00001 for the Immune Epitope Database and the Danish MRC (Medical Research Council) award DFF-6110-00644.

LITERATURE CITED

1. Lund O, Nielsen M, Lundegaard C, Kesmir C, Brunak S. 2005. *Immunological Bioinformatics*. Cambridge, MA: MIT Press
2. Davis MM, Krogsgaard M, Huse M, Huppa J, Lillemeier BF, Li Q. 2007. T cells as a self-referential, sensory organ. *Annu. Rev. Immunol.* 25:681–95
3. Rock KL, Reits E, Neefjes J. 2016. Present yourself! By MHC class I and MHC class II molecules. *Trends Immunol*. 37(11):724–37
4. Kobayashi KS, van den Elsen PJ. 2012. NLRC5: a key regulator of MHC class I-dependent immune responses. *Nat. Rev. Immunol.* 12(12):813–20
5. Wieczorek M, Abualrous ET, Sticht J, Álvaro-Benito M, Stolzenberg S, et al. 2017. Major histocompatibility complex (MHC) class I and MHC class II proteins: conformational plasticity in antigen presentation. *Front. Immunol.* 8:292
6. Thorsby E. 2009. A short history of HLA. *Tissue Antigens* 74(2):101–16
7. Levine BB, Ojeda A, Benacerraf B. 1963. Studies on artificial antigens. III. The genetic control of the immune response to hapten-poly-L-lysine conjugates in guinea pigs. *J. Exp. Med.* 118:953–57
8. McDevitt H. 2002. The discovery of linkage between the MHC and genetic control of the immune response. *Immunol. Rev.* 185:78–85
9. Zinkernagel RM, Doherty PC. 1997. The discovery of MHC restriction. *Immunol. Today* 18(1):14–17
10. Benacerraf B. 1978. A hypothesis to relate the specificity of T lymphocytes and the activity of I region-specific Ir genes in macrophages and B lymphocytes. *J. Immunol.* 120(6):1809–12
11. Rosenthal AS. 1978. Determinant selection and macrophage function in genetic control of the immune response. *Immunol. Rev.* 40:136–52
12. Web of Stories. 2017. Jan Klein: period of confusion in immunology with many false claims. Interview, Aug. 3. **https://www.youtube.com/watch?v=dC7Cy926u_s**
13. Shimonkevitz R, Colon S, Kappler JW, Marrack P, Grey HM. 1984. Antigen recognition by H-2-restricted T cells. II. A tryptic ovalbumin peptide that substitutes for processed antigen. *J. Immunol.* 133(4):2067–74
14. Townsend AR, Gotch FM, Davey J. 1985. Cytotoxic T cells recognize fragments of the influenza nucleoprotein. *Cell* 42(2):457–67
15. Watts TH, Brian AA, Kappler JW, Marrack P, McConnell HM. 1984. Antigen presentation by supported planar membranes containing affinity-purified I-Ad. *PNAS* 81(23):7564–68
16. Dembić Z, Haas W, Weiss S, McCubrey J, Kiefer H, et al. 1986. Transfer of specificity by murine alpha and beta T-cell receptor genes. *Nature* 320(6059):232–38
17. Babbitt BP, Allen PM, Matsueda G, Haber E, Unanue ER. 1985. Binding of immunogenic peptides to Ia histocompatibility molecules. *Nature* 317(6035):359–61
18. Buus S, Sette A, Colon SM, Jenis DM, Grey HM. 1986. Isolation and characterization of antigen-Ia complexes involved in T cell recognition. *Cell* 47(6):1071–77
19. Buus S, Sette A, Colon SM, Miles C, Grey HM. 1987. The relation between major histocompatibility complex (MHC) restriction and the capacity of Ia to bind immunogenic peptides. *Science* 235(4794):1353–58
20. Sette A, Buus S, Colon S, Smith JA, Miles C, Grey HM. 1987. Structural characteristics of an antigen required for its interaction with Ia and recognition by T cells. *Nature* 328(6129):395–99
21. Sette A, Buus S, Colon S, Miles C, Grey HM. 1988. I-Ad-binding peptides derived from unrelated protein antigens share a common structural motif. *J. Immunol.* 141(1):45–48
22. Buus S, Sette A, Grey HM. 1987. The interaction between protein-derived immunogenic peptides and Ia. *Immunol. Rev.* 98:115–41
23. Buus S, Sette A, Colon SM, Grey HM. 1988. Autologous peptides constitutively occupy the antigen binding site on Ia. *Science* 242(4881):1045–47
24. Bjorkman PJ, Saper MA, Samraoui B, Bennett WS, Strominger JL, Wiley DC. 1987. Structure of the human class I histocompatibility antigen, HLA-A2. *Nature* 329(6139):506–12

25. Bjorkman PJ, Saper MA, Samraoui B, Bennett WS, Strominger JL, Wiley DC. 1987. The foreign antigen binding site and T cell recognition regions of class I histocompatibility antigens. *Nature* 329(6139):512–18
26. Brown JH, Jardetzky TS, Gorga JC, Stern LJ, Urban RG, et al. 1993. Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1. *Nature* 364(6432):33–39
27. Yewdell JW, Bennink JR. 1999. Immunodominance in major histocompatibility complex class I–restricted T lymphocyte responses. *Annu. Rev. Immunol.* 17:51–88
28. Sidney J, Southwood S, Moore C, Oseroff C, Pinilla C, et al. 2013. Measurement of MHC/peptide interactions by gel filtration or monoclonal antibody capture. *Curr. Protoc. Immunol.* 100:18.3.1–18.3.36
29. Pedersen LØ, Nissen MH, Hansen NJ, Nielsen LL, Lauenmøller SL, et al. 2001. Efficient assembly of recombinant major histocompatibility complex class I molecules with preformed disulfide bonds. *Eur. J. Immunol.* 31(10):2986–96
30. Stryhn A, Pedersen LØ, Romme T, Holm CB, Holm A, Buus S. 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(8):1911–18
31. Vita R, Overton JA, Greenbaum JA, Ponomarenko J, Clark JD, et al. 2015. The Immune Epitope Database (IEDB) 3.0. *Nucleic Acids Res.* 43:D405–12
32. Falk K, Rotzschke O, Stevanovic S, Jung G, Rammensee HG. 1991. Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. *Nature* 351(6324):290–96
33. Engelhard VH. 1994. Structure of peptides associated with class I and class II MHC molecules. *Annu. Rev. Immunol.* 12:181–207
34. Nielsen M, Lund O, Buus S, Lundegaard C. 2010. MHC class II epitope predictive algorithms. *Immunology* 130(3):319–28
35. Lundegaard C, Lund O, Buus S, Nielsen M. 2010. Major histocompatibility complex class I binding predictions as a tool in epitope discovery. *Immunology* 130(3):309–18
36. Mei S, Li F, Leier A, Marquez-Lago TT, Giam K, et al. 2019. A comprehensive review and performance evaluation of bioinformatics tools for HLA class I peptide-binding prediction. *Brief. Bioinform.* 2019:bbz051
37. Andreatta M, Nielsen M. 2016. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics* 32(4):511–17
38. Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M. 2017. NetMHCpan-4.0: improved peptide–MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. *J. Immunol.* 199(9):3360–68
39. Bassani-Sternberg M, Gfeller D. 2016. Unsupervised HLA peptidome deconvolution improves ligand prediction accuracy and predicts cooperative effects in peptide–HLA interactions. *J. Immunol.* 197(6):2492–99
40. O'Donnell TJ, Rubinsteyn A, Bonsack M, Riemer AB, Laserson U, Hammerbacher J. 2018. MHCflurry: open-source class I MHC binding affinity prediction. *Cell Syst.* 7(1):129–32.e4
41. Parker KC, Bednarek MA, Coligan JE. 1994. Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. *J. Immunol.* 152(1):163–75
42. Rammensee H, Bachmann J, Emmerich NP, Bachor OA, Stevanovic S. 1999. SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics* 50:213–19
43. Peters B, Sette A. 2005. Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method. *BMC Bioinform.* 6:132
44. Jensen KK, Andreatta M, Marcatili P, Buus S, Greenbaum JA, et al. 2018. Improved methods for predicting peptide binding affinity to MHC class II molecules. *Immunology* 154(3):394–406
45. Sette A, Buus S, Appella E, Smith JA, Chesnut R, et al. 1989. Prediction of major histocompatibility complex binding regions of protein antigens by sequence pattern analysis. *PNAS* 86(9):3296–300
46. Rammensee H, Bachmann J, Emmerich NP, Bachor OA, Stevanović S. 1999. SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics* 50(3–4):213–19
47. Bui HH, Sidney J, Peters B, Sathiamurthy M, Sinichi A, et al. 2005. Automated generation and evaluation of specific MHC binding predictive tools: ARB matrix applications. *Immunogenetics* 57(5):304–14

48. Odunsi K, Ganesan T. 2001. Motif analysis of HLA class II molecules that determine the HPV associated risk of cervical carcinogenesis. *Int. J. Mol. Med.* 8(4):405–12
49. Hammer J, Bono E, Gallazzi F, Belunis C, Nagy Z, Sinigaglia F. 1994. Precise prediction of major histocompatibility complex class II-peptide interaction based on peptide side chain scanning. *J. Exp. Med.* 180(6):2353–58
50. Gulukota K, Sidney J, Sette A, DeLisi C. 1997. Two complementary methods for predicting peptides binding major histocompatibility complex molecules. *J. Mol. Biol.* 267(5):1258–67
51. Milik M, Sauer D, Brunmark AP, Yuan L, Vitiello A, et al. 1998. Application of an artificial neural network to predict specific class I MHC binding peptide sequences. *Nat. Biotechnol.* 16(8):753–56
52. Adams HP, Koziol JA. 1995. Prediction of binding to MHC class I molecules. *J. Immunol. Methods* 185(2):181–90
53. Brusic V, Rudy G, Honeyman G, Hammer J, Harrison L. 1998. Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network. *Bioinformatics* 14(2):121–30
54. Buus S, Lauemoller SL, Worning P, Kesmir C, Frimurer T, et al. 2003. Sensitive quantitative predictions of peptide-MHC binding by a "query by committee" artificial neural network approach. *Tissue Antigens* 62(5):378–84
55. Nielsen M, Lundegaard C, Worning P, Lauemoller SL, Lamberth K, et al. 2003. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci.* 12(5):1007–17
56. Zhang C, Bickis MG, Wu F-X, Kusalik AJ. 2006. Optimally-connected hidden Markov models for predicting MHC-binding peptides. *J. Bioinform. Comput. Biol.* 4(5):959–80
57. Mamitsuka H. 1998. Predicting peptides that bind to MHC molecules using supervised learning of hidden Markov models. *Proteins* 33(4):460–74
58. Doytchinova IA, Flower DR. 2001. Toward the quantitative prediction of T-cell epitopes: coMFA and coMSIA studies of peptides with affinity for the class I MHC molecule HLA-A*0201. *J. Med. Chem.* 44(22):3572–81
59. Bhasin M, Singh H, Raghava GPS. 2003. MHCBN: a comprehensive database of MHC binding and non-binding peptides. *Bioinformatics* 19(5):665–66
60. Brusic V, Rudy G, Harrison LC. 1994. MHCPEP: a database of MHC-binding peptides. *Nucleic Acids Res.* 22(17):3663–65
61. Peters B, Tong W, Sidney J, Sette A, Weng Z. 2003. Examining the independent binding assumption for binding of peptide epitopes to MHC-I molecules. *Bioinformatics* 19(14):1765–72
62. Donnes P, Elofsson A. 2002. Prediction of MHC class I binding peptides, using SVMHC. *BMC Bioinform.* 3:25
63. Liu W, Meng X, Xu Q, Flower DR, Li T. 2006. Quantitative prediction of mouse class I MHC peptide binding affinity using support vector machine regression (SVR) models. *BMC Bioinform.* 7:182
64. Peters B, Bui HH, Frankild S, Nielson M, Lundegaard C, et al. 2006. A community resource benchmarking predictions of peptide binding to MHC-I molecules. *PLOS Comput. Biol.* 2(6):e65
65. Madden DR. 1995. The three-dimensional structure of peptide-MHC complexes. *Annu. Rev. Immunol.* 13:587–622
66. Zhang H, Wang P, Papangelopoulos N, Xu Y, Sette A, et al. 2010. Limitations of ab initio predictions of peptide binding to MHC class II molecules. *PLOS ONE* 5(2):e9272
67. Lin HH, Ray S, Tongchusak S, Reinherz EL, Brusic V. 2008. Evaluation of MHC class I peptide binding prediction servers: applications for vaccine research. *BMC Immunol.* 9:8
68. Brown JH, Jardetzky TS, Gorga JC, Stern LJ, Urban RG, et al. 2015. Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1. *J. Immunol* 194(1):5–11
69. Noguchi H, Kato R, Hanai T, Matsubara Y, Honda H, et al. 2002. Hidden Markov model-based prediction of antigenic peptides that interact with MHC class II molecules. *J. Biosci. Bioeng.* 94(3):264–70
70. Salomon J, Flower DR. 2006. Predicting class II MHC-peptide binding: a kernel based approach using similarity scores. *BMC Bioinform.* 7:501
71. Cui J, Han LY, Lin HH, Zhang HL, Tang ZQ, et al. 2007. Prediction of MHC-binding peptides of flexible lengths from sequence-derived structural and physicochemical properties. *Mol. Immunol.* 44(5):866–77

72. Nielsen M, Lundegaard C, Worning P, Hvid CS, Lamberth K, et al. 2004. Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach. *Bioinformatics* 20(9):1388–97
73. Nielsen M, Lund O. 2009. NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction. *BMC Bioinform.* 10:296
74. Nielsen M, Andreatta M. 2017. NNAlign: a platform to construct and evaluate artificial neural network models of receptor-ligand interactions. *Nucleic Acids Res.* 45(W1):W344–49
75. Lin HH, Zhang GL, Tongchusak S, Reinherz EL, Brusic V. 2008. Evaluation of MHC-II peptide binding prediction servers: applications for vaccine research. *BMC Bioinform.* 9(Suppl. 12):S22
76. Karosiene E, Rasmussen M, Blicher T, Lund O, Buus S, Nielsen M. 2013. NetMHCIIpan-3.0, a common pan-specific MHC class II prediction method including all three human MHC class II isotypes, HLA-DR, HLA-DP and HLA-DQ. *Immunogenetics* 65(10):711–24
77. Nielsen M, Justesen S, Lund O, Lundegaard C, Buus S. 2010. NetMHCIIpan-2.0—improved pan-specific HLA-DR predictions using a novel concurrent alignment and weight optimization training procedure. *Immunome Res.* 6:9
78. Lundegaard C, Lund O, Nielsen M. 2008. 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* 24(11):1397–98
79. Gfeller D, Guillaume P, Michaux J, Pak H-S, Daniel RT, et al. 2018. The length distribution and multiple specificity of naturally presented HLA-I ligands. *J. Immunol.* 201(12):3705–16
80. Trolle T, McMurtrey CP, Sidney J, Bardet W, Osborn SC, et al. 2016. The length distribution of class I-restricted T cell epitopes is determined by both peptide supply and MHC allele-specific binding preference. *J. Immunol.* 196(4):1480–87
81. Guillaume P, Picaud S, Baumgaertner P, Montandon N, Schmidt J, et al. 2018. The C-terminal extension landscape of naturally presented HLA-I ligands. *PNAS* 115(20):5083–88
82. McMurtrey C, Trolle T, Sansom T, Remesh SG, Kaever T, et al. 2016. *Toxoplasma gondii* peptide ligands open the gate of the HLA class I binding groove. *eLife* 5:e12556
83. Pymm P, Illing PT, Ramarathinam SH, O'Connor GM, Hughes VA, et al. 2017. MHC-I peptides get out of the groove and enable a novel mechanism of HIV-1 escape. *Nat. Struct. Mol. Biol.* 24(4):387–94
84. Zhao W, Sher X. 2018. Systematically benchmarking peptide-MHC binding predictors: from synthetic to naturally processed epitopes. *PLOS Comput. Biol.* 14(11):e1006457
85. Alvarez B, Reynisson B, Barra C, Buus S, Ternette N, et al. 2019. NNAlign_MA; MHC peptidome deconvolution for accurate MHC binding motif characterization and improved T cell epitope predictions. *Mol. Cell. Proteom.* 18(12):2459–77
86. Paul S, Croft NP, Purcell AW, Tscharke DC, Sette A, et al. 2019. Benchmarking predictions of MHC class I restricted T cell epitopes. bioRxiv 694539. **https://doi.org/10.1101/694539**
87. Bugembe DL, Ekii AO, Ndembi N, Sewanga J, Kaleebu P, Pala P. 2020. Computational MHC-I epitope predictor identifies 95% of experimentally mapped HIV-1 clade A and D epitopes in a Ugandan cohort. *BMC Infect. Dis.* 20(1):172
88. Sette A, Vitiello A, Reherman B, Fowler P, Nayersina R, et al. 1994. The relationship between class I binding affinity and immunogenicity of potential cytotoxic T cell epitopes. *J. Immunol.* 153(12):5586–92
89. Southwood S, Sidney J, Kondo A, del Guercio MF, Appella E, et al. 1998. Several common HLA-DR types share largely overlapping peptide binding repertoires. *J. Immunol.* 160(7):3363–73
90. Rao X, Costa AI, van Baarle D, Kesmir C. 2009. A comparative study of HLA binding affinity and ligand diversity: implications for generating immunodominant $CD8^+$ T cell responses. *J. Immunol.* 182(3):1526–32
91. Schellens IMM, Hoof I, Meiring HD, Spijkers SNM, Poelen MCM, et al. 2015. Comprehensive analysis of the naturally processed peptide repertoire: differences between HLA-A and B in the immunopeptidome. *PLOS ONE* 10(9):e0136417
92. Paul S, Weiskopf D, Angelo MA, Sidney J, Peters B, Sette A. 2013. HLA class I alleles are associated with peptide-binding repertoires of different size, affinity, and immunogenicity. *J. Immunol.* 191(12):5831–39

93. Stranzl T, Larsen MV, Lundegaard C, Nielsen M. 2010. NetCTLpan: pan-specific MHC class I pathway epitope predictions. *Immunogenetics* 62(6):357–68
94. Nielsen M, Andreatta M. 2016. NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. *Genome Med.* 8(1):33
95. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SG. 2015. The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res.* 43:D423–31
96. Sette A, Sidney J. 1999. Nine major HLA class I supertypes account for the vast preponderance of HLA-A and -B polymorphism. *Immunogenetics* 50(3–4):201–12
97. Lund O, Nielsen M, Kesmir C, Petersen AG, Lundegaard C, et al. 2004. Definition of supertypes for HLA molecules using clustering of specificity matrices. *Immunogenetics* 55(12):797–810
98. Nielsen M, Lundegaard C, Blicher T, Lamberth K, Harndahl M, et al. 2007. NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. *PLOS ONE* 2(8):e796
99. Zhang GL, Khan AM, Srinivasan KN, August JT, Brusic V. 2005. MULTIPRED: a computational system for prediction of promiscuous HLA binding peptides. *Nucleic Acids Res.* 33:W172–79
100. Jojic N, Reyes-Gomez M, Heckerman D, Kadie C, Schueler-Furman O. 2006. Learning MHC I–peptide binding. *Bioinformatics* 22(14):e227–35
101. Jacob L, Vert JP. 2008. Efficient peptide–MHC-I binding prediction for alleles with few known binders. *Bioinformatics* 24(3):358–66
102. Zhang H, Lund O, Nielsen M. 2009. The PickPocket method for predicting binding specificities for receptors based on receptor pocket similarities: application to MHC-peptide binding. *Bioinformatics* 25(10):1293–99
103. Sturniolo T, Bono E, Ding J, Raddrizzani L, Tuereci O, et al. 1999. Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. *Nat. Biotechnol.* 17(6):555–61
104. Bordner AJ, Mittelmann HD. 2010. MultiRTA: a simple yet reliable method for predicting peptide binding affinities for multiple class II MHC allotypes. *BMC Bioinform.* 11:482
105. Pfeifer N, Kohlbacher O. 2008. Multiple instance learning allows MHC class II epitope predictions across alleles. In *Proceedings of the 8th International Workshop on Algorithms in Bioinformatics*, ed. KA Crandall, J Lagergren, pp. 210–21. Berlin: Springer
106. Nielsen M, Lundegaard C, Blicher T, Peters B, Sette A, et al. 2008. Quantitative predictions of peptide binding to any HLA-DR molecule of known sequence: NetMHCIIpan. *PLOS Comput. Biol.* 4(7):e1000107
107. Zhang L, Udaka K, Mamitsuka H, Zhu S. 2012. Toward more accurate pan-specific MHC-peptide binding prediction: a review of current methods and tools. *Brief. Bioinform.* 13(3):350–64
108. Trolle T, Metushi IG, Greenbaum JA, Kim Y, Sidney J, et al. 2015. Automated benchmarking of peptide-MHC class I binding predictions. *Bioinformatics* 31(13):2174–81
109. Andreatta M, Trolle T, Yan Z, Greenbaum JA, Peters B, Nielsen M. 2018. An automated benchmarking platform for MHC class II binding prediction methods. *Bioinformatics* 34(9):1522–28
110. van der Burg SH, Visseren MJ, Brandt RM, Kast WM, Melief CJ. 1996. Immunogenicity of peptides bound to MHC class I molecules depends on the MHC-peptide complex stability. *J. Immunol.* 156(9):3308–14
111. Harndahl M, Rasmussen M, Roder G, Dalgaard Pedersen I, Sørensen M, et al. 2012. Peptide-MHC class I stability is a better predictor than peptide affinity of CTL immunogenicity. *Eur. J. Immunol.* 42(6):1405–16
112. Harndahl M, Rasmussen M, Roder G, Buus S. 2011. Real-time, high-throughput measurements of peptide-MHC-I dissociation using a scintillation proximity assay. *J. Immunol. Methods* 374(1–2):5–12
113. Jørgensen KW, Rasmussen M, Buus S, Nielsen M. 2013. NetMHCstab—predicting stability of peptide:MHC-I complexes; impacts for CTL epitope discovery. *Immunology* 141:18–26
114. Rasmussen M, Fenoy E, Harndahl M, Kristensen AB, Nielsen IK, et al. 2016. Pan-specific prediction of peptide-MHC class I complex stability, a correlate of T cell immunogenicity. *J. Immunol.* 197(4):1517–24

115. Nielsen M, Lundegaard C, Lund O, Kesmir C. 2005. The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage. *Immunogenetics* 57(1–2):33–41
116. Eggers M, Boes-Fabian B, Ruppert T, Kloetzel PM, Koszinowski UH. 1995. The cleavage preference of the proteasome governs the yield of antigenic peptides. *J. Exp. Med.* 182(6):1865–70
117. Peters B, Bulik S, Tampe R, Van Endert PM, Holzhütter H-G. 2003. Identifying MHC class I epitopes by predicting the TAP transport efficiency of epitope precursors. *J. Immunol.* 171(4):1741–49
118. Bhasin M, Raghava GPS. 2004. Analysis and prediction of affinity of TAP binding peptides using cascade SVM. *Protein Sci*. 13(3):596–607
119. Larsen MV, Lundegaard C, Lamberth K, Buus S, Brunak S, 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(8):2295–303
120. Tenzer S, Peters B, Bulik S, Schoor O, Lemmel C, 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(9):1025–37
121. Nussbaum AK, Dick TP, Keilholz W, Schirle M, Stevanović S, et al. 1998. Cleavage motifs of the yeast 20S proteasome β subunits deduced from digests of enolase 1. *PNAS* 95(21):12504–9
122. Toes RE, Nussbaum AK, Degermann S, Schirle M, Emmerich NP, et al. 2001. Discrete cleavage motifs of constitutive and immunoproteasomes revealed by quantitative analysis of cleavage products. *J. Exp. Med.* 194(1):1–12
123. Chang S-C, Momburg F, Bhutani N, Goldberg AL. 2005. The ER aminopeptidase, ERAP1, trims precursors to lengths of MHC class I peptides by a "molecular ruler" mechanism. *PNAS* 102(47):17107–12
124. Saveanu L, Carroll O, Lindo V, Del Val M, Lopez D, et al. 2005. Concerted peptide trimming by human ERAP1 and ERAP2 aminopeptidase complexes in the endoplasmic reticulum. *Nat. Immunol.* 6(7):689–97
125. Caron E, Kowalewski DJ, Chiek Koh C, Sturm T, Schuster H, Aebersold R. 2015. Analysis of major histocompatibility complex (MHC) immunopeptidomes using mass spectrometry. *Mol. Cell. Proteom.* 14(12):3105–17
126. Abelin JG, Keskin DB, Sarkizova S, Hartigan CR, Zhang W, et al. 2017. Mass spectrometry profiling of HLA-associated peptidomes in mono-allelic cells enables more accurate epitope prediction. *Immunity* 46(2):315–26
127. Prilliman K, Lindsey M, Zuo Y, Jackson KW, Zhang Y, Hildebrand W. 1997. Large-scale production of class I bound peptides: assigning a signature to HLA-B*1501. *Immunogenetics* 45(6):379–85
128. Andreatta M, Lund O, Nielsen M. 2013. Simultaneous alignment and clustering of peptide data using a Gibbs sampling approach. *Bioinformatics* 29(1):8–14
129. Bassani-Sternberg M, Pletscher-Frankild S, Jensen LJ, Mann M. 2015. Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation. *Mol. Cell Proteom.* 14(3):658–73
130. Nielsen M, Connelley T, Ternette N. 2018. Improved prediction of bovine leucocyte antigens (BoLA) presented ligands by use of mass-spectrometry-determined ligand and in vitro binding data. *J. Proteome Res.* 17(1):559–67
131. Pearson H, Daouda T, Granados DP, Durette C, Bonneil E, et al. 2016. MHC class I-associated peptides derive from selective regions of the human genome. *J. Clin. Investig.* 126(12):4690–701
132. Bassani-Sternberg M, Chong C, Guillaume P, Solleder M, Pak H, et al. 2017. Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allostery regulating HLA specificity. *PLOS Comput. Biol.* 13(8):e1005725
133. Barra C, Alvarez B, Paul S, Sette A, Peters B, et al. 2018. Footprints of antigen processing boost MHC class II natural ligand predictions. *Genome Med*. 10(1):84
134. Kaufman J. 2018. Generalists and specialists: a new view of how MHC class I molecules fight infectious pathogens. *Trends Immunol.* 39(5):367–79
135. Svitek N, Hansen AM, Steinaa L, Saya R, Awino E, et al. 2014. Use of "one-pot, mix-and-read" peptide-MHC class I tetramers and predictive algorithms to improve detection of cytotoxic T lymphocyte responses in cattle. *Vet. Res.* 45(1):50

136. Andreatta M, Jurtz VI, Kaever T, Sette A, Peters B, Nielsen M. 2017. Machine learning reveals a non-canonical mode of peptide binding to MHC class II molecules. *Immunology* 152(2):255–64
137. Liepe J, Marino F, Sidney J, Jeko A, Bunting DE, et al. 2016. A large fraction of HLA class I ligands are proteasome-generated spliced peptides. *Science* 354(6310):354–58
138. Faridi P, Li C, Ramarathinam SH, Vivian JP, Illing PT, et al. 2018. A subset of HLA-I peptides are not genomically templated: evidence for cis- and trans-spliced peptide ligands. *Sci. Immunol.* 3(28):eaar3947
139. Peters B, Brenner SE, Wang E, Slonim D, Kann MG. 2018. Putting benchmarks in their rightful place: the heart of computational biology. *PLOS Comput. Biol.* 14(11):e1006494
140. Braendstrup P, Mortensen BK, Justesen S, Osterby T, Rasmussen M, et al. 2014. Identification and HLA-tetramer-validation of human $CD4^+$ and $CD8^+$ T cell responses against HCMV proteins IE1 and IE2. *PLOS ONE* 9(4):e94892
141. Perez CL, Larsen MV, Gustafsson R, Norstrom MM, Atlas A, et al. 2008. Broadly immunogenic HLA class I supertype-restricted elite CTL epitopes recognized in a diverse population infected with different HIV-1 subtypes. *J. Immunol.* 180(7):5092–100
142. Weiskopf D, Angelo MA, de Azeredo EL, Sidney J, Greenbaum JA, et al. 2013. Comprehensive analysis of dengue virus-specific responses supports an HLA-linked protective role for $CD8^+$ T cells. *PNAS* 110(22):E2046–53
143. Lindestam Arlehamn CS, Sette A. 2014. Definition of CD4 immunosignatures associated with MTB. *Front. Immunol.* 5:124
144. Bjerregaard A-M, Nielsen M, Jurtz V, Barra CM, Hadrup SR, et al. 2017. An analysis of natural T cell responses to predicted tumor neoepitopes. *Front. Immunol.* 8:1566
145. Editorial. 2017. The problem with neoantigen prediction. *Nat. Biotechnol.* 35(2):97
146. Hoof I, van Baarle D, Hildebrand WH, Keşmir C. 2012. Proteome sampling by the HLA class I antigen processing pathway. *PLOS Comput. Biol.* 8(5):e1002517
147. Juncker AS, Larsen MV, Weinhold N, Nielsen M, Brunak S, Lund O. 2009. Systematic characterisation of cellular localisation and expression profiles of proteins containing MHC ligands. *PLOS ONE* 4(10):e7448
148. Yadav M, Jhunjhunwala S, Phung QT, Lupardus P, Tanguay J, et al. 2014. Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. *Nature* 515(7528):572–76
149. Croft NP, Smith SA, Pickering J, Sidney J, Peters B, et al. 2019. Most viral peptides displayed by class I MHC on infected cells are immunogenic. *PNAS* 116(8):3112–17
150. Marino F, Chong C, Michaux J, Bassani-Sternberg M. 2019. High-throughput, fast, and sensitive immunopeptidomics sample processing for mass spectrometry. In *Immune Checkpoint Blockade: Methods and Protocols*, ed. Y Pico de Coaña, pp. 67–79. New York: Humana
151. Andreatta M, Nicastri A, Peng X, Hancock G, Dorrell L, et al. 2019. MS-rescue: a computational pipeline to increase the quality and yield of immunopeptidomics experiments. *Proteomics* 19(4):e1800357
152. Konda P, Murphy JP, Nielsen M, Gujar S. 2019. Enhancing mass spectrometry-based MHC-I peptide identification through a targeted database search approach. In *Immunoproteomics: Methods and Protocols*, ed. KM Fulton, SM Twine, pp. 301–7. New York: Humana
153. Bassani-Sternberg M. 2018. Mass spectrometry based immunopeptidomics for the discovery of cancer neoantigens. In *Peptidomics: Methods and Strategies*, ed. M Schrader, L Fricker, pp. 209–21. New York: Humana
154. Trolle T, Nielsen M. 2014. NetTepi: an integrated method for the prediction of T cell epitopes. *Immunogenetics* 66(7–8):449–56
155. Dhanda SK, Karosiene E, Edwards L, Grifoni A, Paul S, et al. 2018. Predicting HLA CD4 immunogenicity in human populations. *Front. Immunol.* 9:1369
156. Frankild S, de Boer RJ, Lund O, Nielsen M, Kesmir C. 2008. Amino acid similarity accounts for T cell cross-reactivity and for "holes" in the T cell repertoire. *PLOS ONE* 3(3):e1831
157. Bresciani A, Paul S, Schommer N, Dillon MB, Bancroft T, et al. 2016. T-cell recognition is shaped by epitope sequence conservation in the host proteome and microbiome. *Immunology* 148(1):34–39
158. Paul S, Sidney J, Sette A, Peters B. 2016. TepiTool: a pipeline for computational prediction of T cell epitope candidates. *Curr. Protoc. Immunol.* 114:18.19.1–18.19.24

159. Schubert B, Lund O, Nielsen M. 2013. Evaluation of peptide selection approaches for epitope-based vaccine design. *Tissue Antigens* 82(4):243–51
160. Buus S. 1999. Description and prediction of peptide-MHC binding: the "human MHC project." *Curr. Opin. Immunol.* 11(2):209–13
161. Sidney J, Becart S, Zhou M, Duffy K, Lindvall M, et al. 2017. Citrullination only infrequently impacts peptide binding to HLA class II MHC. *PLOS ONE* 12(5):e0177140
162. Zarling AL, Polefrone JM, Evans AM, Mikesh LM, Shabanowitz J, et al. 2006. Identification of class I MHC-associated phosphopeptides as targets for cancer immunotherapy. *PNAS* 103(40):14889–94
163. Andersen MH, Bonfill JE, Neisig A, Arsequell G, Sondergaard I, et al. 1999. Phosphorylated peptides can be transported by TAP molecules, presented by class I MHC molecules, and recognized by phosphopeptide-specific CTL. *J. Immunol.* 163(7):3812–18
164. Klausen MS, Jespersen MC, Nielsen H, Jensen KK, Jurtz VI, et al. 2019. NetSurfP-2.0: improved prediction of protein structural features by integrated deep learning. *Proteins* 87(6):520–27
165. Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, et al. 2019. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* 37(4):420–23
166. Jurtz VI, Johansen AR, Nielsen M, Almagro Armenteros JJ, Nielsen H, et al. 2017. An introduction to deep learning on biological sequence data: examples and solutions. *Bioinformatics* 33(22):3685–90
167. Bulik-Sullivan B, Busby J, Palmer CD, Davis MJ, Murphy T, et al. 2018. Deep learning using tumor HLA peptide mass spectrometry datasets improves neoantigen identification. *Nat. Biotechnol.* 37:56–63
168. Cho S, Kang J. 2019. Dissociation kinetics of TAPBPR-MHC class I complex. *Mol. Immunol.* 114:661–62
169. Dash P, Fiore-Gartland AJ, Hertz T, Wang GC, Sharma S, et al. 2017. Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature* 547(7661):89–93
170. Glanville J, Huang H, Nau A, Hatton O, Wagar LE, et al. 2017. Identifying specificity groups in the T cell receptor repertoire. *Nature* 547(7661):94–98
171. Lanzarotti E, Marcatili P, Nielsen M. 2019. T-cell receptor cognate target prediction based on paired α and β chain sequence and structural CDR loop similarities. *Front. Immunol.* 10:2080
172. Lanzarotti E, Marcatili P, Nielsen M. 2017. Identification of the cognate peptide-MHC target of T cell receptors using molecular modeling and force field scoring. *Mol. Immunol.* 94:91–97
173. Jurtz VI, Jessen LE, Bentzen AK, Jespersen MC, Mahajan S, et al. 2018. NetTCR: sequence-based prediction of TCR binding to peptide-MHC complexes using convolutional neural networks. bioRxiv 433706. **https://doi.org/10.1101/433706**
174. Zhang S-Q, Ma K-Y, Schonnesen AA, Zhang M, He C, et al. 2018. High-throughput determination of the antigen specificities of T cell receptors in single cells. *Nat. Biotechnol.* 36:1156–59
175. Flesch IEA, Woo W-P, Wang Y, Panchanathan V, Wong Y-C, et al. 2010. Altered $CD8^+$ T cell immunodominance after vaccinia virus infection and the naive repertoire in inbred and F_1 mice. *J. Immunol.* 184(1):45–55
176. Castelli FA, Szely N, Olivain A, Casartelli N, Grygar C, et al. 2013. Hierarchy of CD4 T cell epitopes of the ANRS Lipo5 synthetic vaccine relies on the frequencies of pre-existing peptide-specific T cells in healthy donors. *J. Immunol.* 190(11):5757–63

Annual Review of Biomedical Data Science

Volume 3, 2020

Contents

Errata

An online log of corrections to *Annual Review of Biomedical Data Science* articles may be found at http://www.annualreviews.org/errata/biodatasci