MAPS: Analyzing Peptide Binding Subsites in Major Histocompatibility Complexes

Jinbu Wang
Lehigh University
Bethlehem, PA

Brian Y. Chen
Lehigh University
Bethlehem, PA
chen@cse.lehigh.edu

ABSTRACT

The adaptive immune system is a defense system against repeated infection. In order to trigger the immune response, antigen peptides from the infecting agent must first be recognized by the Major Histocompatibility Complex (MHC) proteins. Identifying peptides that bind to MHC class II is thus a critical step in vaccine development. We hypothesize that comparing individual subsites of the peptide binding groove could predict the individual amino acids of possible antigens. This modularized approach to individual subsites could reduce the amount of training data needed for accurate classification while also reducing computing times associated with molecular simulation and docking. To test this hypothesis, we evaluated the capability of two classification techniques and multiple modular representations of the MHC subsites to correctly classify the binding preference categories of P1 subsites of MHC class II structures. Our results shows that the average accuracies are 0.87 for K-mean and 0.95 for SVM with all feature vector configurations. Our results demonstrate that accurate predictions on individual binding subsites is possible, pointing to larger scale applications predicting whole-peptide preferences.

ACM Reference Format:
Jinbu Wang and Brian Y. Chen. 2018. MAPS: Analyzing Peptide Binding Subsites in Major Histocompatibility Complexes. In *9th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (ACM-BCB'18), August 29-September 1, 2018, Washington, DC, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3233547.3233710

1 INTRODUCTION

The T-cell-mediated immune response relies on the proteins of the Major Histocompatibility Complex (MHC), which exhibit two main categories: class I and class II. Class II proteins can be found on the surface of T-cells [5, 12], where they recognize antigen peptides, which are protein fragments left behind by infecting agents. Once an antigen is recognized, the T-cell can trigger the adaptive immune response against that agent. For this reason, determining the peptides that are recognized by a given MHC II protein is a critical step in vaccine design [13, 19, 31]: Rather than using antigens from weakened agents, patients can be sensitized to synthetic peptides, which are easier to produce and safer because they are disconnected from the infective agent. This paper develops MAPS (MHC Analysis of Peptide Subsites), a novel structural analysis technique intended to assist in predicting the antigen binding preferences of a given class II complex.

The peptide recognition site of all class II complexes consists of two interrupted helices and eight antiparallel β-sheets [32]. Instead of packing against each other, the two helices form a binding groove (fig. 1) that accommodates the peptide in an extended conformation. Within the binding groove, there are five cavities, or 'pockets', called P1, P4, P6, P7 and P9. The pockets in these positions accommodate the first, fourth, sixth, seventh, and nineth amino acids of the peptide. While each pocket is somewhat promiscuous in the range of amino acids it accepts, the collected binding preferences of all pockets largely determines which peptides bind any MHC in class II. P1, frequently called an anchor residue binding pocket, is the largest pocket and the major determinant of peptide binding to MHC II [14, 32]. For this reason, we focus our initial evaluation of MAPS on predicting ranges of P1 binding preferences.

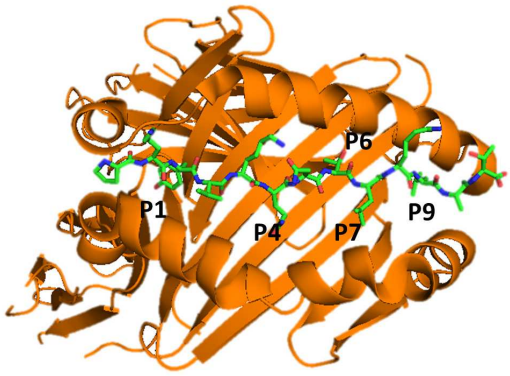

Figure 1: Binding groove and bounded peptide of 1DLH MHC protein. P1,P4,P6,P7 and P9 are the bounding pockets facing the MHC II in the binding groove.

MAPS is trained on pockets with different amino acid preferences at the same position. Given a new pocket, MAPS analyzes the geometry of its molecular surface and electrostatic isopotentials and then predicts which amino acids the given pocket prefers to bind. We represent the molecular surface and the electrostatic isopotentials of each pocket as geometric solids. This approach follows earlier work, where we observed that comparisons of geometric solids can distinguish steric [9] and electrostatic [8] differences

ACM-BCB'18, August 29-September 1, 2018, Washington, DC, USA

ACM ISBN 978-1-4503-5794-4/18/08…$15.00
https://doi.org/10.1145/3233547.3233710

that influence specificity. The geometric solids are then used to train a classifier that makes predictions on query pockets.

This paper evaluates two crucial design questions about MAPS. First, we assess which molecular characteristics can be used to achieve the highest classification sensitivity and specificity. Not only do we consider the molecular surface of the pocket, but because different degrees of electric charge may best reveal differences in binding specificity, we also consider multiple positive and negative electrostatic isopotentials, used together or separately. We also evaluate classification approaches using k means clustering and support vector machines (SVMs) to predict the binding preferences of the pocket. These possibilities are evaluated in detail in our results.

We constructed our data set with 86 peptide-MHC class II complexes out of the available 115 MHC class II related structures from the Protein Data Bank (PDB) [34]. Our data set includes most of peptide-MHC class structures. These structures were selected specifically because we were able to identify experimental results documenting the binding preferences of the P1 pocket in each complex. On these data, we exhaustively cross-validated MAPS to assess the accuracy of the classifiers and pocket representations above. Our results point to a new strategy for predicting the peptide binding preferences of individual MHCs.

2 RELATED WORK

Many methods for predicting peptide binding to MHC molecules can be categorized into two broad categories: sequence-based and structure-based methods. MAPS is a structure-based method but it avoids some design constraints of both categories, illustrating a new approach to the problem.

2.1 Sequence-based methods

Sequence-based methods use motifs and machine learning to predict peptide sequences that bind to a specific MHC. Methods using motifs scan the sequences of binding peptides and build qualitative or quantitative weight matrices that describe the space of compatible peptides [15, 20, 22, 30, 33]. These positional weight matrices contain 9 columns, one for each amino acid position in the peptide. The i_{th} column of a weight matrix has 20 elements, representing the probability of one of the 20 canonical amino acids occuring in the i_{th} position of a peptide that binds in the MHC. Altogether, the matrix defines the peptide binding preferences of the MHC protein, and it is expected that different MHC proteins exhibit different weight matrices. Unfortunately, weight matrices can be inaccurate predictors of binding when the training data is incomplete [11]. In addition, because weight matrices simplify peptide binding into a series of independent amino acid preferences, inaccuracies can also arise from effects that amino acids in the peptide have on their neighbors [27].

Sequence-based machine learning methods predict the binding affinity of MHC II using artificial neural networks (ANNs) [7, 22], hidden Markov models (HMMs) [23] and SVMs [4, 16]. These approaches are generally trained on sets of peptides that bind and those that do not. A few machine learning methods are trained on more detailed categories such as high binders, intermediate binders, weak binders and nonbinders [21]. The average area under the ROC curve (AUC) is 0.871 for various MHC II alleles [18]. The reliability of learning methods also relies on comprehensive training data that outlines the boundaries of the binding preferences of a given MHC. Typically, 50-100 peptide binding measurements are required to build a model with reasonable accuracy for each MHC I [26, 37], but MHC II binding preferences are broader than those of class I.

MAPS differs from these sequence-based methods because it can focus on classifying individual pockets. While this paper focusses on P1 to prove the concept, MAPS can be independently trained to produce predictions for the other pockets. These modular predictions can then be integrated into a picture of MHC preferences for the overall peptide. The advantage of this approach is that the peptide preferences of the MHC can be predicted as long as examples of the pocket preferences exist in the training set, even if the pocket preferences never occur in the same MHC. While this approach does not consider the influences that adjacent amino acids have on each other, it does not require training sets with the combinatorial scale of existing methods.

2.2 Structure based methods

Structure-based methods employ molecular modeling and molecular docking techniques to predict MHC binding preferences. Molecular modeling methods use molecular dynamics simulations [10], monte carlo simulations [22], or ab initio computations [38] to simulate the atomic motion of the MHC II and the peptide during the binding process. After simulation, the binding potential energies of simulated MHC-peptide complex structures are computed using statistical potentials derived from the existing 3D protein structure [1, 6, 29, 38]. These statistical potentials mitigate the need to find experimental data that describes the affinity between a range of peptides and a given MHC, and they also create the freedom to examine new MHCs for which no peptide binding preferences have been established. Unfortunately, the time consuming nature of simulation bottlenecks the construction of a combinatorial training set that examines many possible binding peptides.

Unlike modeling approaches, docking methods assume that both the MHC and peptide are rigid. Specifically, that interatomic distances in MHC II are fixed and that the peptide is rigid. Holding the MHC II fixed, molecular docking methods attempt to rotate and translate the binding peptide into the binding cavity of the MHC II. Using statistical potentials similar to those used by modeling approaches, molecular docking methods search the space of rigid complexes for a complex with lowest potential energy. Because molecular flexibility is not considered, molecular docking methods [2, 3, 24] are faster than molecular modeling methods, but still many orders of magnitude slower than sequence based methods.

In contrast to these methods, MAPS is a structure based method that performs classification by comparing MHC structures rather than simulation or docking. Comparisons have been shown in the past to be able to reveal steric and electrostatic similarities and differences in binding preferences [8, 9], and they can be localized to individual pockets without loss of accuracy, to support the modular approach of MAPS. Structure comparisons can also be performed much more rapidly than simulation or docking, because they do not require the iterative computation of potential energies.

However, unlike other structure-based methods, comparisons depend more on the availability of existing structures, because classifications are performed by simularity and dissimilarity rather than energies derived from first principles. Our results will enable us to assess how accurately structure comparisons can support this difficult classification problem.

3 METHODS

MAPS is composed of 3 steps: First, we generate solid representations of binding pockets using constructive solid geometry (CSG). Second, we generate feature vectors from the binding pockets. Finally, we perform classifier training and prediction.

3.1 Constructive solid geometry

CSG is a category of techniques for building solid objects from other solids with three basic operations: Boolean Union, Boolean intersection and Boolean difference (Figure 2). These operations perform logical operations on geometric solids as if they were mathmatical sets containing points in space. CSG was developed for computer aided design of machine parts [35] and adapted for computer graphics to represent geometric solids [17]. MAPS uses CSG to represent proteins and binding pockets as geometric solids.

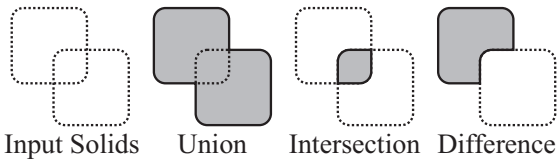

Figure 2: Operations in constructive solid geometry (CSG)

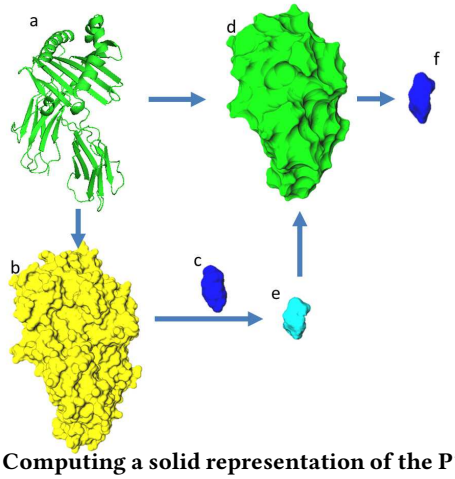

Figure 3: Computing a solid representation of the P1 pocket. (a) Cartoon representation of an MHC II structure. (b) The molecular surface of the MHC. (c) The Boolean union of all P1 amino acids in the data set. (d) The envelope surface. (e) The Boolean difference of the union of P1 amino acids minus the molecular surface. (f) The final P1 pocket: The Boolean intersection between (e) and the envelope surface.

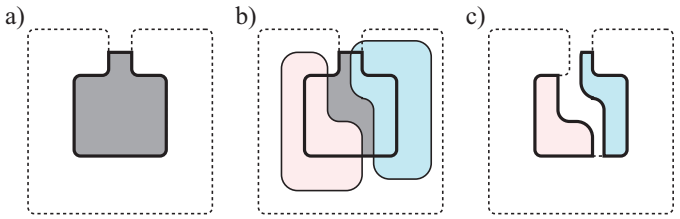

Figure 4: Computing a solid representation of an electrostatic isopotential in the binding pocket. (a) The molecular surface of the MHC (dotted) and the P1 binding pocket (gray, solid outline). (b) Isopotentials of the entire structure (light black outlines). Positive isopotentials are blue, negative isopotentials are red. Solid representations of the electrostatic isopotentials in the binding pocket (red and blue with heavy solid outlines).

3.2 Generating solid representations

MAPS generates solid representations of pockets using CSG operations (Fig. 3). First, using Ska [36], all of the MHC II protein peptide complexes were structurally aligned to HLA-DRB1 (pdb: 4MD5), which was randomly selected. Second, the peptide-MHC complexes were split into MHC II proteins and binding peptides, and the molecular surfaces of MHC II proteins were generated using trollbase [36] (Fig. 3b). Third, we compute the Boolean union of all the molecular surfaces of all the P1 amino acids from the bound peptides in our dataset (Fig. 3c). These amino acids overlap tightly because they are in complex with the aligned MHC structures, and the union of these amino acids loosely defines the location of the P1 pocket in the working surface. Fourth, using trollbase, we compute an *envelope surface* using a 5.0 Å probe (Fig. 3d). This surface defines a boundary between the cavities of the protein and the bulk solvent around it. Fifth, we compute the Boolean difference of the union of P1 amino acids minus the molecular surface (Fig. 3e), identifying the region occupied by possible peptide residues that is also solvent accessible near the MHC structure. Finally, we generate the Boolean intersection between the difference and the envelope surface (Fig. 3f), eliminating any part of the remaining region that is within the bulk solvent and thus not in a pocket. This geometric solid defines the P1 pocket.

We generate solid representations of electrostatic isopotentials following the technique described by VASP-E [8] (Fig. 4). We begin with the solid representation of the P1 pocket (Fig. 4a). Second, using VASP-E, we compute electrostatic isopotentials of the entire protein (Fig. 4b). To evaluate several thresholds, we generated isopotentials at +/-1, +/-3 and +/-5 kT/e. Finally, we compute Boolean intersections between the isopotentials and the pocket (Fig. 4c). The resulting intersections describe solvent accessible regions that are exposed to positively or negatively charged parts of the electrostatic field.

3.3 Generating feature vectors

Once the P1 pockets and their electrostatic isopotentials have been generated, we translate them into feature vectors for classification. First, note that all MHC structures were structurally aligned, so all P1 pockets and their electrostatic isopotentials are also superposed. Second, we generate a lattice of uniformly sized cubes. We refer to the length of the side of one cube as the resolution of this

lattice, which is held at .5 Å in this work. Using Boolean operations and the Surveyor's formula [28], we measure the volume of intersection between each cube and each solid representation, and normalize the intersection volume to the volume of the cube. The normalized volumes inside all cubes in the bounding box, including the cubes that have no intersection, are treated as the elements of a feature vector. We process three kinds of solids in this way: binding cavities, which represent the solvent accessible shape of the binding cavity that accommodates one of the amino acids in the peptide, positive electrostatic isopotentials at a positive potential threshold, and negative electrostatic isopotentials at a negative potential threshold.

Uncertain of which solids might yield the most informative data for classification, we generated feature vectors from binding cavities alone (which we refer to simply as *shape*), from positive electrostatic isopotentials alone, and from negative electrostatic isopotentials alone. The feature vectors that combine multiple such characteristics were generated by concatenating their elements. All combinations of characteristics were tested in our experimentation. In our results, we refer to feature vectors that include binding cavities and positive electrostatic isopotentials as *shape + positive*, those that include all features as *shape + positive + negative*, and so on. Generating feature vectors from solid representations was implemented in Python and C++. Generation time was less than 30 seconds on a 2.4GHz CPU.

3.4 Classification

In order to select a classifier that could effectively identify pockets with similar and different binding preferences, we examined the classification performance achieved by K-means clustering and SVMs, which are suitable for high dimensional feature vectors. K-means clustering is an unsupervised machine learning method with unlabeled data. It finds K groups in the data, where K is an input variable describing the number of clusters. This algorithm works iteratively to classify each input data point to one of K clusters basing on feature similarity, constantly seeking to minimize the objective function $J = \sum_{j=1}^{k} \sum_{i}^{n} \|x_i^j - c_j\|^2$, where $\|x_i^j - c_j\|^2$ is the Euclidean distance between the j_{th} data vector x^j and the cluster centroid c_j, n is number of data point vectors and k is the number of clusters. SVMs are a class of supervised machine learning methods that use labeled data for training. SVMs are effective in high dimensional data, even when the number of dimensions is greater than the number of samples, as is the case with our data. SVMs use training data to select an objective function from several categories of mathematical kernel functions. We considered SVMs with linear, polynomial, and Gaussian kernels. Our experimentation evaluates the capacities of K-means clustering and SVMs to distinguish P1 pockets with different binding preferences. All classification was automated using the Sci-kit learn[25] python library and default parameters are used in both methods.

All feature vectors of MHC II binding pockets are taken as the input for the classification of binding pockets with the number of clusters $k = 2$. The K-means method was used to classify the binding preferences of binding pocket and predict the binding preference with a higher accuracy. Our experimentation using the K-means and SVM methods employed the Sci-kit learn[25] python library. We evaluated the capacity of the K-means clustering algorithm to separate binding cavities that prefer either charged amino acids or hydrophobic amino acids.

3.5 Data Set

Training our classifiers required ground truth data describing the binding preferences of the P1 pockets in MHC II proteins. To gather this data, we began with the 86 MHC II protein-peptide complexes stored in the PDB as of 1 January 2017. To determine the binding preferences of the P1 pockets in each structure, significant literature search was performed to identify experimental evidence substantiating the binding preferences at each P1. These references are provided in Table 1. Some papers referenced in the PDB only illustrated the MHC structure and did not describe the binding preferences of each binding pocket, leading us to examine related publications. Since classification is performed on the P1 binding pocket, and since some publications examined multiple pockets, only citations for the binding preferences of the pocket are referenced. According to the literature, the majority of P1 structures in our dataset prefer hydrophobic residues, and approximately 10 percent prefer charged residues. For this reason, the K-means and SVM classifiers described above were used to perform dichotomous classification.

Variations in the way binding preferences are described in publication led us to define some preferences as "nonpolar" rather than "hydrophobic" in Table 1. For the purposes of classification, we treat them as members of the same category. Likewise, all pockets that prefer charged amino acids prefer negatively charged amino acids, except one, which prefers positively charged amino acids, but we treat them as members of the same category. In future work, subdividing or expanding these categories is a logical extension of this proof of concept.

3.6 Accuracy, Sensitivity, and Specificity

Since our classifiers are performing dichotomous classification between "hydrophobic" and "charged" categories, we measure prediction accuracy by counting the number of predictions. Normally, dichotomous prediction scenarios involve predicting whether or not a particular statement is true, but in this case the two categories are biophysical opposites in nature. As such, we count true positives (TPs), false positives (FPs), true negatives (TNs) and false negatives (FNs) as follows:

- TP=# MHC II predicted to prefer hydrophobic amino acids in P1, and actually do
- FP=# MHC II predicted to prefer hydrophobic amino acids in P1, and actually prefer charged
- TN=# MHC II predicted to prefer charged residues in P1, and actually do
- FN=# MHC II predicted to prefer charged residues in P1, and actually prefer hydrophobic

To establish best practices during calibration, we compute accuracy as a single-dimensional value to illustrate the relative performance of our classifiers in different configurations. We evaluate accuracy as $\frac{TP+TN}{TP+TN+FP+FN}$. Accuracy is the fraction of correct predictions relative to the total number of predictions made, and it ranges from 0.0 (least accurate) to 1.0 (most accurate). Once configured, to provide detailed information on the performance of our

Table 1: P1 pocket binding preferences of MHC II

#	pdb	Res.	Pref.	#	pdb	Res.	Pref.	#	pdb	Res.	Pref.	#	pdb	Res.	Pref.
1	1DLH	Y	H	2	1A6A	Y	H	3	2SEB	M	H	4	1AQD	W	H
5	1IAK	D	C	6	1BX2	V	H	7	1FYT	Y	H	8	1F3J	R	C
9	1FV1	F	H	10	1JK8	E	C	11	1HQR	F	H	12	1HXY	Y	H
13	1J8H	Y	H	14	1H15	Y	H	15	1KG0	W	H	16	1LO5	W	H
17	1KLG	I	H	18	1KLU	Y	H	19	1JWM	Y	H	20	1JWS	Y	H
21	1JWU	Y	H	22	1PYW	F	H	23	1S9V	P	H	24	1UVQ	L	H
25	1T5W	Y	H	26	1T5X	Y	H	27	1SJE	V	H	28	1SJH	V	H
29	1R5I	W	H	30	1YMM	V	H	31	1ZGL	F	H	32	2G9H	Y	H
33	2FSE	F	H	34	2IAM	I	H	35	2IAN	I	H	36	2NNA	E	C
37	2OJE	W	H	38	2Q6W	W	H	39	2ICW	Y	H	40	3C5J	I	H
41	3LQZ	F	H	42	3PDO	M	H	43	3PGC	M	H	44	3PGD	M	H
45	3L6F	Y	H	46	2XN9	Y	H	47	3MBE	R	H	48	4Z7W	E	C
49	3QXA	M	H	50	3QXD	M	H	51	3S4S	Y	H	52	3S5L	Y	H
53	4GG6	E	C	54	4GBX	Y	H	55	4AEN	M	H	56	4AH2	M	H
57	4D8P	E	C	58	4H1L	I	H	59	4H25	I	H	60	4H26	I	H
61	4IS6	L	H	62	4MCY	V	H	63	4MCZ	Y	H	64	4MD0	W	H
65	4MD4	W	H	66	4MD5	V	H	67	4MDI	V	H	68	4MDJ	V	H
69	4I5B	V	H	70	4P2O	L	H	71	4P2Q	L	H	72	4P2R	V	H
73	4P4K	F	H	74	4P4R	F	H	75	4P5K	F	H	76	4P5M	W	H
77	4P57	Ph	H	78	3WEX	K	C	79	4OZF	P	H	80	4OZG	P	H
81	4OZH	P	H	82	4OZI	P	H	83	4C56	W	H	84	4OV5	A	H
85	4Z7U	E	C	86	4Z7V	E	C								

Accuracy (positive kT/e, negative kT/e)	(1,-1)	(1,-3)	(1,-5)	(3,-1)	(3,-3)	(3,-5)	(5,-1)	(5,-3)	(5,-5)
Shape	0.793	0.793	0.793	0.793	0.793	0.793	0.793	0.793	0.793
Positive	0.953	0.953	0.953	0.953	0.953	0.953	0.953	0.953	0.953
Negative	0.867	0.847	0.706	0.867	0.847	0.706	0.867	0.847	0.706
Shape + Positive	0.793	0.793	0.793	0.805	0.805	0.805	0.805	0.805	0.805
Shape + Negative	0.863	0.854	0.720	0.900	0.854	0.720	0.900	0.854	0.720
Positive + Negative	0.964	0.847	0.705	0.952	0.847	0.706	0.940	0.847	0.706
Shape + Positive + Negative	0.900	0.850	0.720	0.900	0.854	0.720	0.900	0.854	0.720

Table 2: Prediction Accuracy of K-means clustering at several electrostatic isopotential thresholds.

Accuracy (positive kT/e, negative kT/e)	(1,-1)	(1,-3)	(1,-5)	(3,-1)	(3,-3)	(3,-5)	(5,-1)	(5,-3)	(5,-5)
Shape	0.927	0.927	0.927	0.927	0.927	0.927	0.927	0.927	0.927
Positive	0.976	0.976	0.976	0.965	0.965	0.965	0.965	0.965	0.965
Negative	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Shape + Positive	0.927	0.927	0.927	0.915	0.915	0.915	0.927	0.927	0.927
Shape + Negative	0.950	0.927	0.927	0.950	0.927	0.927	0.950	0.927	0.927
Positive + Negative	0.928	0.929	0.941	0.928	0.941	0.941	0.940	0.929	0.918
Shape + Positive + Negative	0.938	0.939	0.927	0.938	0.927	0.915	0.938	0.939	0.915

Table 3: Prediction Accuracy of SVMs at several electrostatic isopotential thresholds

classifiers, we compute sensitivity and specificity. In our classification scheme, sensitivity of hydrophobic amino acids denotes the probability of all hydrophobic binding preferences that that are predicted as hydrophobic binding preferences and specificity of hydrophobic amino acids denotes the probability of predicted hydrophobic binding preferences that are actual hydrophobic binding preferences. The sensitivity and specificity of hydrophobic amino acids are computed $\frac{TP}{TP+FN}$ and $\frac{TP}{TP+FP}$, respectively. The sensitivity and specificity are evaluated as $\frac{TN}{TN+FP}$ and $\frac{TN}{TN+FN}$ respectively.

4 RESULTS

Our results evaluate the prediction accuracy of K-means and SVM-based classifiers, as well as the contributions of different steric and electrostatic representations of the P1 pocket as a feature vector. We also evaluated the accuracy of predictions made with SVM classifiers using different kernel functions. In each case, exhaustive leave-one-out validation was used to ensure that results reflect the true accuracy, sensitivity, or specificity of each classifier.

4.1 Calibrating electrostatic isopotentials in MAPS

Electrostatic isopotentials can be computed at multiple thresholds. Considering the possibility that both K-means and SVM-based classifiers might be more accurate when using electrostatic isopotentials at informative thresholds, we trained them with feature vectors that included up to two isopotentials, one positively charged and one negatively charged. We evaluated the classification accuracy of both classifiers with thresholds equal to -5.0, -3.0, -1.0, 1.0, 3.0 and 5.0 kT/e. We also considered feature vectors that incorporate molecular surfaces only ("shape"), which was simply the geometry of P1 without any electrostatic isopotentials, and thus identical at all charge thresholds. Also, we considered feature vectors that incorporate only one of the two isopotentials, which are referred to as "Positive" and "Negative" in Tables 2 and 3. In each case, accuracy was computed as the number of correct predictions divided by the totally number of predictions made, and the pairs of isopotential thresholds were described as an ordered pair of values, (positive threshold, negative threshold).

Overall, feature vectors with positive and negative isopotentials at 1.0 kT/e and -1.0 kT/e respectively produced the most accurate predictions on average.

Feature vector type	RBF	linear	Polynomial
Shape	0.878	0.915	0.878
Positive	0.882	0.965	0.882
Negative	0.880	1.00	0.880
Shape + Positive	0.875	0.927	0.875
Shape + Negative	0.878	0.938	0.878
Negative + Positive	0.880	0.916	0.880
Shape + Postive + Negative	0.875	0.938	0.875

Table 4: SVM Accuracy on three kernels

4.2 Selecting SVM Kernels

SVMs can be trained with a variety of kernel functions. We evaluated three kernels: Gaussian kernels, also known as radial basis functions (RBF), linear kernels, and polynomial kernels. We evaluated classifications performed with these kernels on the feature vectors we created, including positive and negative isopotentials alone, and molecular surfaces only. Following the observations of Section 4.1, electrostatic isopotentials were generated at 1.0 kT/e and -1.0 kT/e to produce maximum accuracy. These results are summarized in Table 4.

Overall, linear SVM kernels exhibited average accuracy on all feature vectors equal to .943, which was distinctly more accurate than gaussian kernels (.878) and polynomial kernels (.878). Evaluated on every feature vector, linear kernels outperformed RBF and polynomial kernels.

4.3 Classifier Sensitivity and Specificity

Having established that both a K-means classifier and an SVM classifier perform best with feature vectors that include positive and negative isopotentials at 1.0 kT/e and -1.0 kT/e, and having observed that the SVM classifier performs best with a linear kernel, we measured the sensitivity and specificity of both our calibrated predictors. We also measured the sensitivity and specificity of the other feature vector configurations that we examined earlier. Results on K-means clustering are outlined in Table 5, and Table 6 describes results on SVMs.

Overall, linear SVMs performed with sensitivity and specificity greater than the K-means classifier. The sensitivity and specificity of prediction on hydrophobic pockets was nearly perfect for most feature vector configurations, with Negative feature vectors performing best. Sensitivity and specificity on P1 pockets that prefer charged residues was generally similar between K-means and SVM classifiers, while classification sensitivity on pockets that prefer hydrophobic residues was lower when using K-means classification on most feature vector configurations.

5 DISCUSSION

Our results with MAPS have evaluated several classifiers for predicting the binding preferences of the P1 binding pocket of MHC class II structures. The performance of our most successful classifiers demonstrate proof of concept that it is possible to use structure comparison to accurately predict the binding preferences of a pocket in the MHC binding groove. This validate our hypothesis that comparing individual subsites of the binding groove can predict preferences for individual amino acids. Due to the modularity of this comparison strategy, these results point to future applications of this strategy for predicting the binding preferences at each pocket in the peptide binding groove. Together, modular predictions could support the discovery of new antigens that could be used to protect patients from infective agents.

Beyond modularity, the advantage of this approach is a seamless capacity to consider a diverse range of biophysical phenomena without the computational expense of molecular modeling and docking. MAPS demonstrates that high resolution steric and electrostatic descriptors can be produced with a volumetric representation and integrated into the classification system. This result indicates that future opportunities exist for including additional biophysical descriptors that can be compared in the same volumetric manner.

REFERENCES

[1] S. Aldulaijan and J. A. Platts. 2010. Theoretical prediction of a peptide binding to major histocompatibility complex II. *J Mol Graph Model* 29, 2 (2010), 240–5.

[2] M Atanasova, I Dimitrov, DR Flower, and I Doytchinova. 2011. MHC class II binding prediction by molecular docking. *Molecular Informatics* 30, 4 (2011), 368–375.

[3] M. Atanasova, A. Patronov, I. Dimitrov, D. R. Flower, and I. Doytchinova. 2013. EpiDOCK: a molecular docking-based tool for MHC class II binding prediction. *Protein Eng Des Sel* 26, 10 (2013), 631–4.

[4] M. Bhasin and G. P. Raghava. 2004. SVM based method for predicting HLA-DRB1*0401 binding peptides in an antigen sequence. *Bioinformatics* 20, 3 (2004), 421–3.

[5] PJ Bjorkman, MA Saper, B Samraoui, WS Bennett, JL Strominger, and DC Wiley. 1987. The foreign antigen binding site and T cell recognition regions. *Nature* 329 (1987), 512–518.

[6] A. J. Bordner. 2010. Towards universal structure-based prediction of class II MHC epitopes for diverse allotypes. *PLoS One* 5, 12 (2010), e14383.

[7] V. Brusic, G. Rudy, G. Honeyman, J. Hammer, and L. Harrison. 1998. Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network. *Bioinformatics* 14, 2 (1998), 121–30.

[8] B. Y. Chen. 2014. VASP-E: specificity annotation with a volumetric analysis of electrostatic isopotentials. *PLoS Comput Biol* 10, 8 (2014), 1–17.

[9] Brian Y Chen and Barry Honig. 2010. VASP: a volumetric analysis of surface properties yields insights into protein-ligand binding specificity. *PLoS Comput Biol* 6, 8 (2010), 11.

CATEGORY	Accuracy	Sens(hydro)	Spec(hydro)	Sens(charge)	Spec(charge)
Shape	0.79	0.79	0.97	0.80	0.34
Positive	0.95	1.0	0.95	0.60	1.0
Negative	0.87	0.92	0.93	0.50	0.45
Shape + Positive	0.79	0.81	0.95	0.70	0.33
Shape + Negative	0.86	0.93	0.96	0.70	0.58
Positive + Negative	0.96	1.0	0.96	0.70	1.0
Shape + Positive + Negative	0.90	0.93	0.96	0.70	0.58

Table 5: Sensitivity and Specificity of P1 preference predictions using K-means, using +1/-1 kT/e

CATEGORY	Accuracy	Sens(hydro.)	Spec(hydro)	Sens(charge)	Spec(charge)
Shape	0.93	0.96	0.96	0.70	0.70
Positive	0.98	1.00	0.97	0.80	1.00
Negative	1.00	1.00	1.00	1.00	1.00
Shape + Positive	0.93	0.96	0.96	0.70	0.70
Shape + Negative	0.95	0.99	0.96	0.70	0.88
Positive + Negative	0.93	0.96	0.96	0.70	0.70
Shape + Positive + Negative	0.94	0.97	0.96	0.70	0.78

Table 6: Sensitivity and Specificity of P1 preference predictions using Linear SVM

[10] I. Doytchinova, P. Petkov, I. Dimitrov, M. Atanasova, and D. R. Flower. 2011. HLA-DP2 binding prediction by molecular dynamics simulations. *Protein Sci* 20, 11 (2011), 1918–28.

[11] Y. El-Manzalawy, D. Dobbs, and V. Honavar. 2008. On evaluating MHC-II binding peptide prediction methods. *PLoS One* 3, 9 (2008), e3268.

[12] K. C. Garcia and E. J. Adams. 2005. How the T cell receptor sees antigen - A structural view. *Cell* 122, 3 (2005), 333–336.

[13] Stephen J Goodswen, Paul J Kennedy, and John T Ellis. 2014. Enhancing In Silico Protein-Based Vaccine Discovery for Eukaryotic Pathogens Using Predicted Peptide-MHC Binding and Peptide Conservation Scores. *PloS one* 9, 12 (2014), e115745.

[14] Juergen Hammer, Charles Belunis, David Bolin, Joanne Papadopoulos, Robert Walsky, Jacqueline Higelin, Waleed Danho, Francesco Sinigaglia, and Zoltan A Nagy. 1994. High-affinity binding of short peptides to major histocompatibility complex class II molecules by anchor combinations. *Proceedings of the National Academy of Sciences* 91, 10 (1994), 4456–4460.

[15] J. Hammer, E. Bono, F. Gallazzi, C. Belunis, Z. Nagy, and F. Sinigaglia. 1994. Precise prediction of major histocompatibility complex class II-peptide interaction based on peptide side chain scanning. *J Exp Med* 180, 6 (1994), 2353–8.

[16] Jing Huang and Feng Shi. 2005. *Prediction of MHC class II epitopes using Fourier analysis and support vector machines.* Springer, 21–30.

[17] Tao Ju, Frank Losasso, Scott Schaefer, and Joe Warren. 2002. Dual contouring of hermite data. In *ACM transactions on graphics (TOG)*, Vol. 21. ACM, 339–346.

[18] E. Karosiene, M. Rasmussen, T. Blicher, O. Lund, S. Buus, and M. Nielsen. 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 (2013), 711–24. https://doi.org/10.1007/s00251-013-0720-y

[19] H. H. Lin, G. L. Zhang, S. Tongchusak, E. L. Reinherz, and V. Brusic. 2008. Evaluation of MHC-II peptide binding prediction servers: applications for vaccine research. *BMC Bioinformatics* 9 Suppl 12 (2008), S22.

[20] Keith W Marshall, K Jeff Wilson, James Liang, Andrea Woods, Dennis Zaller, and Jonathan B Rothbard. 1995. Prediction of peptide affinity to HLA DRB1* 0401. *The journal of immunology* 154, 11 (1995), 5927–5933.

[21] M. Nielsen, O. Lund, S. Buus, and C. Lundegaard. 2010. MHC class II epitope predictive algorithms. *Immunology* 130, 3 (2010), 319–28.

[22] M. Nielsen, C. Lundegaard, P. Worning, C. S. Hvid, K. Lamberth, S. Buus, S. Brunak, and O. Lund. 2004. Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach. *Bioinformatics* 20, 9 (2004), 1388–97.

[23] H. Noguchi, R. Kato, T. Hanai, Y. Matsubara, H. Honda, V. Brusic, and T. Kobayashi. 2002. Hidden Markov model-based prediction of antigenic peptides that interact with MHC class II molecules. *J Biosci Bioeng* 94, 3 (2002), 264–70.

[24] A. Patronov, I. Dimitrov, D. R. Flower, and I. Doytchinova. 2011. Peptide binding prediction for the human class II MHC allele HLA-DP2: a molecular docking approach. *BMC Struct Biol* 11 (2011), 32.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[26] Bjoern Peters and Alessandro Sette. 2005. Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method. *BMC bioinformatics* 6, 1 (2005), 132.

[27] Björn Peters, Weiwei Tong, John Sidney, Alessandro Sette, and Zhiping Weng. 2003. Examining the independent binding assumption for binding of peptide epitopes to MHC-I molecules. *Bioinformatics* 19, 14 (2003), 1765–1772.

[28] J Schaer and MG Stone. 1991. *Face traverses and a volume algorithm for polyhedra.* Springer, 290–297.

[29] H. D. Schafroth and C. A. Floudas. 2004. Predicting peptide binding to MHC pockets via molecular modeling, implicit solvation, and global optimization. *Proteins* 54, 3 (2004), 534–56.

[30] Harpreet Singh and GPS Raghava. 2001. ProPred: prediction of HLA-DR binding sites. *Bioinformatics* 17, 12 (2001), 1236–1237.

[31] Satarudra Prakash Singh and Bhartendu Nath Mishra. 2012. Prediction model of MHC Class-II binding peptide motifs using sequence weighting method for vaccine design. In *Advances in Computing and Communications (ICACC), 2012 International Conference on.* IEEE, 234–237.

[32] L. J. Stern, J. H. Brown, T. S. Jardetzky, J. C. Gorga, R. G. Urban, J. L. Strominger, and D. C. Wiley. 1994. Crystal structure of the human class II MHC protein HLA-DR1 complexed with an influenza virus peptide. *Nature* 368, 6468 (1994), 215–221.

[33] Tiziana Sturniolo, Elisa Bono, Jiayi Ding, Laura Raddrizzani, Oezlem Tuereci, Ugur Sahin, Michael Braxenthaler, Fabio Gallazzi, Maria Pia Protti, and Francesco Sinigaglia. 1999. Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. *Nature biotechnology* 17, 6 (1999), 555–561.

[34] J. L. Sussman, D. Lin, J. Jiang, N. O. Manning, J. Prilusky, O. Ritter, and E. E. Abola. 1998. Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. *Acta Crystallogr D Biol Crystallogr* 54, Pt 6 Pt 1 (1998), 1078–84.

[35] Herbert B Voelcker and Aristides AG Requicha. 1977. Geometric modeling of mechanical parts and processes. *Computer* 10, 12 (1977), 48–57.

[36] An-Suei Yang and Barry Honig. 2000. An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance. *Journal of molecular biology* 301, 3 (2000), 665–678.

[37] Hao Zhang, Ole Lund, and Morten Nielsen. 2009. The PickPocket method for predicting binding specificities for receptors based on receptor pocket similarities: application to MHC-peptide binding. *Bioinformatics* 25, 10 (2009), 1293–1299.

[38] H. Zhang, P. Wang, N. Papangelopoulos, Y. Xu, A. Sette, P. E. Bourne, O. Lund, J. Ponomarenko, M. Nielsen, and B. Peters. 2010. Limitations of Ab initio predictions of peptide binding to MHC class II molecules. *PLoS One* 5, 2 (2010), e9272.