

Epitope Prediction Algorithms for Peptide-based Vaccine Design

Liliana Florea, Bjarni Halldórsson, Oliver Kohlbacher,
Russell Schwartz, Stephen Hoffman, and Sorin Istrail
Celera/Applied Biosystems
Sorin.Istrail@celera.com

Abstract

Peptide-based vaccines, in which small peptides derived from target proteins (epitopes) are used to provoke an immune reaction, have attracted considerable attention recently as a potential means both of treating infectious diseases and promoting the destruction of cancerous cells by a patient's own immune system. With the availability of large sequence databases and computers fast enough for rapid processing of large numbers of peptides, computer aided design of peptide-based vaccines has emerged as a promising approach to screening among billions of possible immune-active peptides to find those likely to provoke an immune response to a particular cell type. In this paper, we describe the development of three novel classes of methods for the prediction of class I epitopes. Each one of the three classes of methods gives a specific set of insights into the epitope prediction problem. We present a quadratic programming approach that can be trained on quantitative as well as qualitative data. The second method uses linear programming to counteract the fact that our training data contains mostly positive examples. The third class of methods uses sequence profiles obtained by clustering known epitopes to score candidate peptides. By integrating these methods, using a simple voting heuristic, we achieve improved accuracy over the state of the art.

1 Introduction

The vertebrate class I immune system monitors protein expression within cells and induces lysis in those exhibiting aberrant expression. This immune response is crucial to targeting virus-infected cells, which must express non-self peptides during the course of reproducing an infecting virus, and provides one line of defense against cancer, as proteins that are mis-spliced or over-expressed due to genetic damage can trigger an immune response that results in destruction of the damaged cells. Tumor immunotherapy [24] is a recent therapeutic approach aimed at priming the

immune system to respond to malignant cells in patients whose immune systems are not responding on their own, or to boost insufficient responses.

The human immune system detects antigens produced in its own cells (i.e., foreign viral proteins or over-expressed or mutated self-proteins) by means of the MHC class I pathway. In a first step, these antigens are cleaved in the cytosol of the cell to produce individual peptides, which are then transported into the endoplasmic reticulum (ER) by the transporter associated with antigen processing (TAP), where some of them bind to proteins of the major histocompatibility complex (MHC) and are presented at the cell surface. Recognition of the MHC-peptide complex by the T cell receptor (TCR) found on the surface of cytotoxic T lymphocytes (CTLs) then triggers cytolysis of the peptide-presenting cell (i.e. the tumor cell). One way to boost the immune response towards a specific antigen is thus the administration of peptides derived from this antigen which are recognized by MHC class I. These so-called *class I epitopes* are usually peptides of 8-11 amino acids.

The pathway from protein sequence to vaccine development is lengthy and cost-intensive [7], entailing the development of binding assays for testing the affinity of the selected peptides to the MHC molecules, in vitro assays for measuring the T-cell response, and ultimately in vivo testing of immunogenicity. There is thus considerable incentive to screen candidate peptides computationally prior to assay development, leading to the development of computational methods for detecting immunogenicity. Each of the steps in the antigen-processing pathway adds some specificity to antigen selection. However, the primary and most discriminating point at which sequence specificity constrains antigen recognition appears to be the incorporation of peptides in the MHC complex [17]. Computational methods have therefore concentrated on predicting the MHC binding affinity of candidate peptides. The MHC genes are highly variable among human populations, with different human leukocyte

antigen (HLA) groups (alleles) having different binding specificities, which must be considered by the prediction method.

The MHC class I epitopes bind to a well-defined binding groove on the MHC molecule. The binding mode of the peptides is very specific at the N- and C-termini and somewhat less specific in the middle of the peptides. The main sources of this specificity are the “anchor sites”, which are pockets in the MHC molecule that accommodate certain peptide side chains. Early methods for prediction focused on characterizing likely epitopes by testing for the presence of the appropriate primary anchors [15] and secondary anchor residues [25]. The first approach to large-scale *in silico* epitope prediction based on anchor identification was taken by Rammensee et al. [21], yielding an algorithm for predicting epitopes from protein sequences, and a database (SYFPEITHI; <http://syfpeithi.de/>) of experimentally identified and published motifs. More recently, the pattern-matching algorithm EpiMer [12] was developed based on similar principles, but identifying not just motifs for one allele type, but also MHC “promiscuous” ligands containing patterns that allow binding to more than one type of MHC molecule. Alternatives to these pattern-based methods include statistical and machine learning methods [14, 8] and structure-based methods [23].

One class of prediction methods particularly relevant to the present work is that of matrix-based methods, introduced in Parker et al. [20]. These methods assume that the strength of binding of an epitope to the MHC allele is given by the sum of independent binding contributions of each of the peptide’s amino acids, i.e. the peptide’s binding energy is just the sum of the binding energies of its amino acids at their respective positions. The binding strength of these peptides is experimentally determined either as IC₅₀ values (the concentration of the peptide inhibiting the binding of a reference peptide in half of the bindings sites) or as the half-life of binding for the MHC complex. Both quantities can be related to true binding energies.

A frequently made assumption is that each amino acid a in each position i contributes to the overall binding energy¹ of the peptide independently of the other amino acids. The total binding energy ΔG is thus the sum of the individual contributions g_a^i of the amino acids. The binding energy of a peptide p with nine amino acids $p = a_1 a_2 a_3 a_4 a_5 a_6 a_7 a_8 a_9$ can thus be

¹In fact, these energies are binding *free* energies, but we will use the shorter term binding energy from here on.

modeled as

$$\Delta G(p) = \sum_i g_{a_i}^i. \quad (1)$$

To predict the binding strength of all peptides we now only need to know the binding strength of the 20 possible amino acids in each of the nine binding positions, for a total of 180 values. These 180 values form a 9x20 *binding matrix* \mathbf{B} . Using a binary coding for the sequence, where ones in a 9x20 matrix $\mathbf{S}(p)$ represent the amino acid at the respective position of peptide p , we can easily write $\Delta G(p)$ as

$$\Delta G(p) = \sum_{i,j} \mathbf{B}_{ij} \mathbf{S}_{ij}(p). \quad (2)$$

Most experiments do not yield binding energies directly, but rather binding constants K or the related IC₅₀ values. However, a similar relationship holds for K , based on the individual contributions k_a^i of the amino acids:

$$K = e^{-\frac{RT}{\Delta G}} = \prod_i e^{-\frac{RT}{g_{a_i}^i}} = \prod_i k_{a_i}^i \quad (3)$$

Parker et al. showed how a combination of linear regression and dimensionality reduction (assuming some positions are unimportant or some matrix values are equal) could be used to infer these binding matrices from moderate-sized datasets. A similar approach is taken by EpiMatrix [26].

In this work, we describe three novel classes of methods for predicting MHC binding peptides, and a voting scheme to integrate them into a unified prediction framework that yields improved results over each individual method. Each of the three classes of methods leverages a unique set of insights into the prediction problem. The first two use quadratic programming (QP) and linear programming (LP) techniques, respectively, to derive values for a single weight matrix for each allele. The QP method is akin to the regression approach presented by Parker et al., but borrows from the literature on support vector machines (SVMs) to incorporate semi-quantitative as well as fully qualitative data in building an optimal matrix. In the second method we formulate a linear program to take advantage of the fact that we have more examples of known epitopes than non-epitopes. The third class of methods uses sequence profiles obtained by clustering the known epitopes of a given HLA allele to score candidate peptides. This clustering approach, which assumes that the allele-specific epitopes may belong to several, rather than one, sequence motifs is novel in the epitope prediction literature. The three classes of

methods provide complementary views of the rankings of peptides, by exploring different facets of the training data. To take advantage of their complementarity, while at the same time compensate for their biases, we combine several of these methods using a simple voting scheme into a unified “consensus” method, which has improved prediction accuracy and, in particular, higher recognition rate for moderate binders.

We have tested the performance of our methods for several high-quality benchmark sets of experimentally determined epitopes, for the four HLA alleles occurring most commonly in human populations (A2, A3, A24, B7). In each case, our combined method was competitive or outperformed Parker’s method. In particular, our approaches were better suited to detect moderate-binding peptides, which tend to be the most easily missed by all individual methods.

2 Epitope Prediction Algorithms

In this section we present three classes of algorithms for predicting epitopes based on the patterns learned from examples of known epitopes. The first part describes the Quadratic Programming (QP) approach, the second the Linear Programming (LP) approach, and the third the profile-based approach. Finally, the fourth part describes the voting scheme used to combine the different methods.

2.1 Quadratic Programming - Combining Qualitative and Quantitative Data

The quadratic programming (QP) method was inspired by the literature on support vector machines [6] and in particular, a support vector based method employed by Singh and Kim [28] to predict coiled-coil interactions in protein side chains. The method can be seen as an extension to the linear regression approach taken by Parker et al. [20], but allows us to incorporate information gained from sources alternate to the binding half-life (IC_{50} value) measurements. In their method, the input is a set of peptides x_i and a measurement of each peptide’s binding strengths K_i . We wish to find a vector w of predicted binding constants and an offset c from zero giving a predicted binding constant $x_i^T w - c$ for each a peptide so as to minimize $\sum_i (x_i^T w - c - K_i)^2$; the difference between the predicted and measured binding strengths.

In practice, due to difficulties in calibrating measurements from different experimenters, most of the data available to us is not given in terms of binding strength. We may be given only whether the peptide is an epitope or not, or whether a peptide is a “high,” “low,” or “medium” binder. We can add this data to our model by insisting that the model assigns

epitopes a binding strength greater than the minimum binding strength of known epitopes, that non-epitopes are assigned binding strength less than the minimum binding strength of known epitopes, that the high binders have a higher binding strength than the low and medium binders, and that the medium binders have a higher binding strength than the low binders. Let x_H , x_M , x_L be high, medium, and low binders; x_e be epitopes; x_{ne} be non-epitopes; and IC_{50}^{min} be the minimum binding strength of an epitope. We can then formulate our quadratic program as follows:

$$\begin{aligned} \min_{w,c} \sum_i (x_i^T w - c - K_i)^2 \\ \text{s.t. } x_{H_i}^T w \geq x_{M_j}^T w \quad \forall i, j, \\ x_{H_i}^T w \geq x_{L_j}^T w; \quad x_{M_i}^T w \geq x_{L_j}^T w \quad \forall i, j, \\ x_{e_i}^T w \geq IC_{50}^{min}; \quad x_{ne_i}^T w \leq IC_{50}^{min} \quad \forall i \end{aligned}$$

We note that the number of constraints in this formulation grows quadratically with the number of datapoints. To make the problem more manageable we require that a high binder only be a stronger binder than the average medium binder and the average low binder. Similarly, a medium binder is required to be a stronger binder than the average low binder. Furthermore, the program may have a set of constraints that are not feasible (i.e. no set of parameters w and c will satisfy all the constraints) because of inaccuracies in the experiments and to the inadequacy of the independent binding strength assumption. We therefore penalize violations of the constraints, seeking the set of parameters that are in the least violation to our set of constraints. Letting $\overline{x_M}$, $\overline{x_L}$ be the average medium and low binders and adding the appropriate error terms yields the following final program:

$$\begin{aligned} \min_{w,c} \sum_i (x_i^T w - c - K_i)^2 + \sum_i e_{H_i M}^2 + \sum_i e_{H_i L}^2 \\ + \sum_i e_{M_i L}^2 + \sum_{i,j} e_{e_i}^2 + \sum_{i,j} e_{ne_i}^2 \\ \text{s.t. } x_{H_i}^T w + e_{H_i M} \geq \overline{x_M}^T w \quad \forall i \\ x_{H_i}^T w + e_{H_i L} \geq \overline{x_L}^T w; \quad x_{M_i}^T w + e_{M_i L} \geq \overline{x_L}^T w \quad \forall i \\ x_{e_i}^T w + e_{e_i} \geq IC_{50}^{min}; \quad x_{ne_i}^T w + e_{ne_i} \leq IC_{50}^{min} \quad \forall i \end{aligned}$$

Due to the small number of datapoints (relative to the number of variables), we used several approaches to reduce the space dimensionality (number of variables). These approaches led to different matrices and consequently different predictions. In the first

method, we use data reduction techniques applied in [20] in obtaining their matrices. From these matrices, we derived relations regarding which amino acids are irrelevant for binding and which pairs of amino acids have an equivalent effect on binding. For example, Asp and Glu are assumed equivalent for binding in the first position of epitopes for the HLA-A2 allele. We then incorporated these constraints into the quadratic program. In the second method, instead of working with the amino acids as variables, we use amino acid properties to characterize them. Using two of these properties (hydrophobicity and size) as descriptors, we map the peptide into a $9 \times 2 = 18$ dimensional space. Each amino acid is mapped to a point in two dimensions, where the first coordinate is a measurement of its hydrophobicity, and the second is a measurement of its size. In our third method, we selected five best-fit parameters from the 485 indices reported in the AAindex repository [16] to produce a reduced feature set.

2.2 Linear Programming - Classification Based on Positive Examples

The linear programming formulation was motivated by the fact that our input data is qualitative and consists almost solely of positive examples, or epitopes, as opposed to non-epitopes, which in nature far outnumber epitopes. To correct for this imbalance, we construct artificial binding constants for each amino acid in each position such that epitopes have a high binding strength but the overall sum of the binding constants is minimum. Minimizing the sum of the binding constants can be considered a proxy for minimizing the number of peptides considered epitopes. In order to avoid artificially low objective values we further require that all binding strengths be positive.

More specifically, we construct an artificial binding constants k_a^p for each amino acid in each position p and require each k_a^p to be positive. We further require that the sum of the binding constants for each epitope be greater than 1, an arbitrarily chosen constant. Let e_i be the i -th epitope, e_i^p be the amino acid in position p of epitope e_i . Then we can formulate the optimization as the following linear program:

$$\begin{aligned} \min \quad & \sum_{a \neq sa} \sum_{pos.p} k_a^p \\ \text{s.t.} \quad & \sum_{pos.p} k_{e_i^p}^p \geq 1 \forall i, \\ & k_a^p \geq 0 \forall a, p \end{aligned}$$

2.3 Profile-based Prediction

Profiles, previously introduced in the literature in the context of motif detection in DNA and protein sequences [11, 29], are a natural way to represent motifs in epitope sequences. Indeed, epitope and MHC binding peptides are known to contain allele-specific motifs, in which the sequence composition and degree of sequence variation of the various positions are dictated by the specificity of the secondary structure pocket of the MHC. A sequence profile is a representation of a set of aligned sequences as a table of letter frequencies per column in the alignment, together with a position-specific weight matrix derived from the alignment data. The score for a candidate sequence is then the sum of matrix values for the individual 'letters' in that sequence. The higher the score, the better the candidate sequence conforms to the motif pattern, and the more likely it is that it belongs to that motif class. Our profile-based approach to epitope prediction relies on the following three principles:

- *dimensionality reduction* — Certain amino acids have similar functions at a given position, explained by the compatibility of their side chains with the three-dimensional binding pocket, and therefore can be deemed functionally equivalent. For instance, Leu and Met are both hydrophobic and preferred at the P2 anchor position in HLA-A2 ligands. We divided the amino acids in four classes: hydrophobic=H (Ala, Val, Phe, Pro, Met, Ile and Leu), polar=P (Ser, Thr, Tyr, His, Cys, Asn, Gln, and Trp), charged=C (Asp, Glu, Lys and Arg) and glycine=G (Gly), following the classification of Branden and Tooze [3]. This partitioning assumes a uniform substitution model for amino acids across the peptide and across the range of HLA molecules. The *biochemical signature* of a given peptide or protein sequence is then the concatenation of the individual letter classes; for instance, YLLPCITEV has the signature PH-HHPHPCH.
- *multiple intra-allelic motifs* — We hypothesize that epitopes for any given HLA molecule can be classified in one or *more* motif classes, suggested by the observation among epitopes of a given allele of position dependencies, which point to groups of sequences with distinct features (Figure 1). Two methods (*Aln* and *Ki2*) were developed to divide the sets of biochemical signatures of known epitopes for each HLA allele into clusters corresponding to distinct sequence motifs, each represented by a profile of the aligned

	%	P	C	G	H
(1) uniform	40	20	5	35	
(3) P3	32.2	13.4	11.4	43.0	
(3) (PH) ₅ -restricted P3	38.2	11.8	5.5	44.5	
(4) (GC) ₅ -restricted P3	15.4	17.9	28.2	38.5	

Figure 1: Distributions of amino acid types at P3 in the sequences of HLA-A2 ligands and epitopes: (1) distribution in a uniform distribution model; (2) distribution of residue types at P3 in the original data set; (3) in the subset of sequences that contained a hydrophobic (H) or polar (P) residue at P5; (4) in the subset of sequences with a charged (C) or glycine (G) residue at P5. Note that the amino acid type distribution in the sample set is not random, by comparing lines (1) and (2). Also, that significantly different letter profiles are observed for the sequence subsets restricted to contain PH and CG, respectively, at P5 (lines 3 and 4). Lastly, that C or G type residues at P5 strongly favor glycine (G), and restrict polar (P) amino acids at P3. These observations are based on a data set of 149 samples collected from the SYFPEITHI database (see Results).

sequences.

- *anchor selection* — Anchor positions have a lower degree of variation than the other positions in the peptide. This aspect could not be captured with our amino acid encoding scheme. For instance, a 'hydrophobic' label at P2 of a predicted HLA-A2 epitope can encode any of the seven amino acids Ala, Val, Phe, Pro, Met, Ile, and Leu, all of which will produce the same score, while only five (Ile, Leu, Val, Met and Ala) have been observed in practice, and with widely varying incidence rates. To compensate for the loss of specificity, we use a more differentiated scoring for the combination of amino acids at anchor positions.

For our application, profiles were constructed using the alignment of biochemical signatures of experimentally validated epitopes and ligands. The score of a candidate peptide was computed as the sum of amino acid weights [9]:

$$Score(a_1 a_2 \dots a_k) = \sum_{i=1}^k W_{a_i}$$

with:

$$W_b^i = \log(N_b^i + \frac{1}{2}) - \frac{1}{4} \sum_a \log(N_a^i + \frac{1}{2}),$$

$$a, b \in \{P, H, C, G\}$$

where N_a^i is the number of occurrences of amino acid type a in column i.

To compute a score, each candidate peptide of length 9 (ninemers) is converted into its biochemical signature, which is then aligned with each of the profiles for this HLA allele, and a profile score is computed. The final peptide score is the maximum of individual profile scores. For peptides that are 10 residues long, the scores are computed by taking the average of scores for the seven ninemers obtained by removing exactly one of the non-anchor residues at positions P3, P4, P5, P6, P7, P8 and P9.

We considered two clustering strategies for profile construction: iterative multiple alignment (Aln) and position dependencies reflected by χ^2 tests (Ki2). We also give a separate profile-based method for anchor scoring.

Clustering via iterative multiple sequence alignment (Aln)

The Aln method uses a greedy strategy to start and grow a multiple sequence alignment (profile), starting from a 'seed' pair and optimally choosing the next sequence as the one that maximizes the multiple alignment score. The publicly available program CLUSTALW [31] was used to produce the multiple alignments. At each iteration, sequences that do not improve the current score are removed from the pool, to ensure that the clusters converge towards a unique signal. The procedure stops when no sequences can be recruited. The current cluster is saved as a new profile, and the procedure is repeated with the remaining sequences and a new seed pair. The resulting profiles may include gaps, which are made explicit in the profile model. In the end, some manual intervention may be necessary to redistribute the sequences and constrain the alignment of anchors.

Clustering based on position dependencies (Ki2)

The Ki2 method is based on the assumption that the choice of amino acid at one position in the peptide may influence the distribution of amino acid types on the remaining motif positions (Figure 1). To capture the dependencies between pairs of columns, and between a column and the rest of the alignment positions, we used χ^2 statistical significance tests between the consensus (majority) variable for column i (1 if the amino acid type at that position matches the consensus, 0 otherwise), and the indicator variable for the same column, identifying the amino acid type at that position (P, H, C, G). We use the most significant posi-

tion dependencies to gradually split the set of aligned sequences into disjoint clusters each representing a sequence motif, using a procedure akin to that described in [5].

Given a data set of biochemical epitope signatures, consisting of sequences of equal length, we first assign a consensus amino acid type, or group of types, at each position. For the HLA-A2 epitopes, which have a pronounced proclivity towards hydrophobic and polar residues, the consensus vector is $C = [PH, PH, PH, *, PH, PH, PH, PH, H]$. Then, for each pair of positions (i, j) with $i \neq j$ we compute the $\chi^2_{i,j}$ statistics for C_i versus X_j . We further attempt to select a column by which to divide the set of sequences into two groups: those that contain a consensus letter at that position, and those that do not. Specifically, we choose the column l , if any, with the largest overall association $\chi^2_l = \sum_m \chi^2_{lm}$ with the rest of the positions in the alignment, such that $\chi^2_l \geq K$ (K is a splitting constant that depends on the number of degrees of freedom for the current system), and such that each of the two resulting subsets contains at least five sequences. The procedure is repeated for each of the subsets.

We used the information content of the alignment [27, 30]:

$$I = \sum_{i=1}^9 \sum_{a=P,H,C,G} f_a^i \log(f_a^i / 0.25)$$

to measure the quality of profiles, both before and after a cluster division. In all cases, the information content values of the profiles after the operation were higher than that of the original, showing that the profiles have improved.

Anchor scoring

To account for the increased specificity of amino acid distribution at the anchor positions versus the rest of the positions in the motif we computed an additional profile-based anchor score. A single column profile on the alphabet of all possible amino acids pairs (20x20=400 letters) was constructed from the frequencies of residue pairs at the two designated anchor positions, P2 and C-terminal, in class I epitopes (see Results). Candidate peptides are scored after first stripping the non-anchor positions. The profile score for an anchor is its weight:

$$W_b = \log(N_b + 0.5) - 0.25 \sum_{a=H,P,C,G} \log(N_a + 0.5)$$

We chose an alphabet of residue-pairs to correctly reflect some non-independent pairings of

residues at anchor positions observed from the data. Indeed, the fact that Thr at P2 was only found in combination with Val at P9 among the A2 epitopes, and not with Leu, the other strongly favored amino acid at P9, cannot be expressed in a scoring model in which each of the amino acids in a pair contributes independently to the pair's score.

2.4 Voting - Combining Methods

We combine the predictions from these methods using a simple voting heuristic. For the HLA A2 allele, we use two versions of the quadratic programming method, one using matrices obtained based on the amino acid properties of size and hydrophobicity and one using a dimensionality reduction scheme similar to [20]; the linear programming method; alignment profiles; and anchors. For the other alleles we examined — A3, A24, and B7 — we use Parker's method in combination with our LP, QP, alignment profile, and anchor methods. Our implementation of Parker's method uses the most recent matrices maintained at the NIH Bioinformatics and Molecular Analysis Section (BIMAS) site (http://bimas.dcrf.nih.gov/molbio/hla_bind/). We combine the scores of disparate methods into a single prediction by linearly scaling the score of each method so that it ranges from 0 to 1, then summing the scaled scores of all methods for each candidate peptide. This combination of methods was chosen by comparing the predictions of individual methods, and combinations of different methods presented here, on the set of epitopes from the Influenza proteins reported in [10].

3 Results

We trained our algorithms on the four most commonly occurring HLA alleles — A2, A3, A24 and B7 — with A2 being by far the most common and best studied. For this allele a total of 694 nonamer epitopes were extracted from the MHCPEP database (<http://wehih.wehi.edu.au/mhcpep/>; [4]), of which 359 were annotated with their binding strength categories (high, medium or low). In addition, a set of 101 HLA-A2 epitopes, together with their IC_{50} values, was extracted from [20]. Due to the lack of experimentally determined binding strength data, the QP method could not be applied to the other HLA alleles considered. For the LP method, allele specific nonamer data were extracted from the same MHCPEP database: 118 for A3, 23 for A24, and 56 for B7. For both the QP and LP methods, initial predictions were done for nonamers; tenmer scores were then computed by ignoring the seventh position in the sequence [20]. The

resulting quadratic and linear programs were solve using Matlab [19].

For the profile-based methods, a total of 206 epitope and ligand sequences previously published for the HLA-A2 allele were extracted from the SYFPEITHI database (<http://syfpeithi.de/>; [21]). In addition, data for the HLA-A3 (253 sequences), A24 (148) and B7 (189) alleles were extracted from the MHCPEP database. After eliminating duplicates and sequences with more or less than nine residues from the set of ligands and epitopes for each allele, the remaining distinct ninemers were selected for profile construction. These sets consisted of 146 epitopes for A2, 165 for A3, 104 for A24, and 138 for B7. For anchor scoring, the entire pool of peptides for a given allele, regardless of length, was analyzed to determine the frequencies of amino-acid pairs at the P2 and C-terminal positions.

We used *sensitivity curves* to measure and compare the performance of the epitope prediction methods, on a benchmark of known epitopes. The sensitivity curve plots the sensitivity $S_n(x) = TP/(TP+FN)$ achieved by the method when the top ranking $x\%$ of the peptides are selected and classified as epitopes, where TP and FN are the number of true positive, and false negative examples, respectively, in the reference set. In other words, a sensitivity curve plots the percentage f of epitopes that are found in the top ranking $x\%$ of the peptides, and the values along the x -axis indicate what percentage of the peptides needs to be sequenced and tested in order to obtain some fraction $f\%$ of the epitopes in the pool.

We benchmark our algorithms on three different publicly available reference sets of known epitopes and MHC ligand sequences. In all three cases, we compare our predictions to those produced with an improved version of the matrix-based approach presented in [20], available from the NIH BIMAS site (http://bimas.dcrf.nih.gov/molbio/hla_bind). We have focused on comparisons with this method because of its wide acceptance and frequent use as a reference method (e.g., in [18, 8]) and the availability of raw matrices suitable for use in high-throughput analysis. Our comparisons show that our method performs competitively or better than the improved matrix-based approach (BIMAS). In addition, it achieves significantly better results than the original method presented in [20] (data not shown).

According to a recent survey by Lauemøller et al. [17], prediction methods based on simple motif searches can only identify one out of four binders, while extended motifs, which include the most important primary, secondary and disfavored residues,

identify three out of four binders, albeit at the cost of sequencing 8% of the candidate peptides in the protein [21].

The first reference set was a collection of epitopes and MHC ligands for the HLA-A2 allele discovered in Influenza proteins, reported in [10], consisting of 32 sequences from 10 proteins. The A2 allele is the most common and best studied allele. In addition, it is the only one of the alleles examined for which sufficient quantitative binding data were available to use our QP method. Only one of the 32 sequences was used in training all of the prediction methods.

The second benchmark was based on a set of HLA-B7 epitopes from the West Nile virus polyprotein (GenBank Accession: AF196835) previously characterized by De Groot et al. [13], consisting of 12 experimentally verified sequences. None of these sequences were used in training our methods.

The third benchmark consisted of HLA-A2 epitopes isolated from human cancers. Cancer epitopes would be expected to show some selective biases, since they are generally expressed in self-proteins, and therefore would have to escape immune tolerance in addition to the other selectivity constraints presented on epitopes from foreign antigens. We began with a collection of epitopes reported by Renkvist et al. [22] that were isolated from human tumor cells. We screened out those epitopes that do not occur in the wild-type versions of their respective proteins, by retaining only those epitopes that could be found in the sequences of the Celera predicted proteins (translations of the Celera predicted gene sequences [32], or in the Genbank [2], SwissProt/TrEMBL [1], or PIR databases. This dataset contains 50 sequences, of which 13 were used in training of the LP and QP methods, and 27 were included in the profile training set.

For all of the benchmarks, the sensitivity curves (Figure 2A, B and C) show that the differences in performance between our method and the Parker method are small, with each method performing better on certain sections of the epitope range. Parker's method appears to be more selective in predicting the top ranking epitopes, while our combined method performs consistently better when more complete sets are sought. This implies that while the Parker method is better at locating the strongest binders, our method is superior at identifying those binders that are placed towards the mid and lower ranges of the binding affinity scale. This property is particularly desirable for developing cancer vaccines, in which high-binding self-peptides may lead to host tolerance, and therefore be less effective in inducing an immune response. Fur-

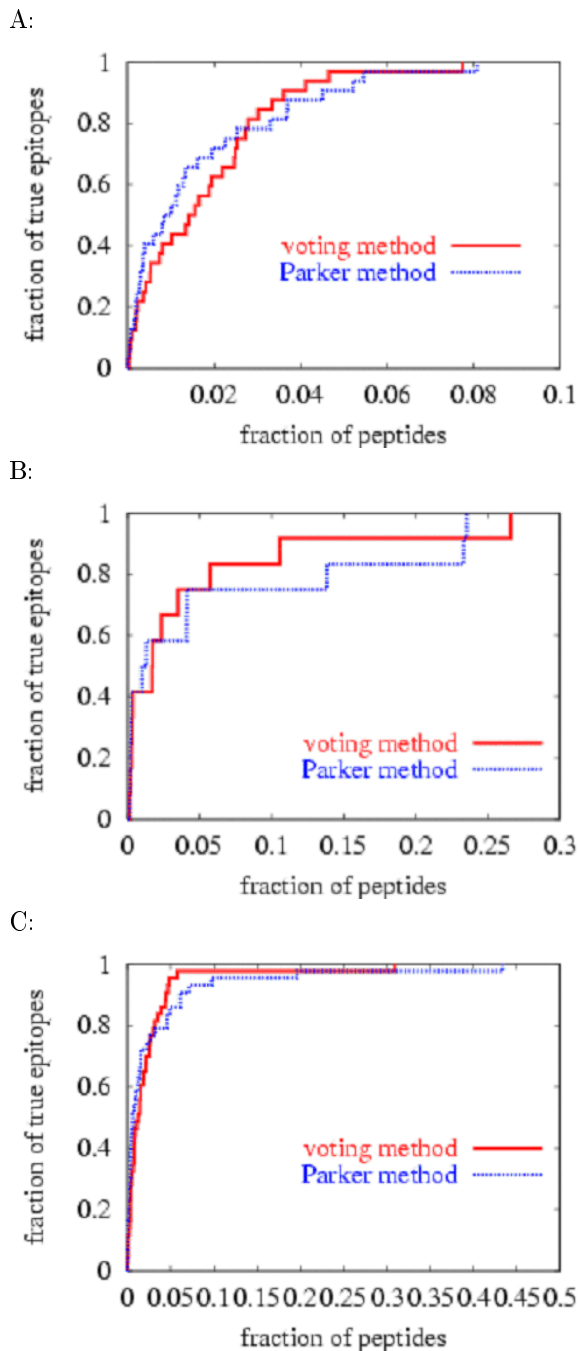


Figure 2: Sensitivity curves for our voting method and for Parker et al.'s method on the test data sets. The fraction f of true epitopes among those in the reference set which are found in the top ranking $x\%$ of the peptides is plotted against the fraction $x\%$ of the peptides selected, for each method. (A) influenza A2 benchmark. (B) West Nile B7 benchmark (C) cancer A2 benchmark

thermore, the ability to efficiently predict a range of peptides containing as complete a set of epitopes as possible given reasonable sequencing costs would strongly benefit high-throughput vaccine development efforts.

4 Discussion

We present several novel methods for MHC epitope prediction. We describe one class of methods based on quadratic programming, one based on linear programming, and several based on profile methods for motif detection. These methods are then combined using a voting scheme to improve performance. Benchmarks developed from literature datasets show that our combined prediction method is competitive with the leading prediction methods reported in the literature. Improved separation of true epitopes from non-immunogenic peptides is achieved both for a dataset of proteins from a foreign pathogen and for a dataset of human proteins known to have triggered immune responses to cancerous tumors. Furthermore, while our methods achieve results overall comparable to Parker's reference method, our method appears better at finding the true MHC binders that are farthest from the established sequence patterns and thus are hardest to detect by any sequence-based method. These are presumably the weakest binders among the true epitopes, and therefore the least likely to promote immune tolerance, and thus the best candidates for cancer vaccine development.

The major limitation on the accuracy of the prediction methods applied in the present work is the lack of available training data. The amount of data available to researchers can be expected to grow as time passes. For the foreseeable future better computational modeling of the problem can be considered to be important for success. One computational strategy that can be pursued is to attempt to incorporate additional biological constraints into the computational learning models, such as those gained from looking at structure, in order to reduce the space of models being examined and thus reduce data dependence. Conversely, better results might be achieved by relaxing existing constraints, such as the assumption of independent residue contributions.

Rapid and accurate high-throughput screening of epitopes can be instrumental for accelerated development of vaccines against novel strains of known pathogens or, in conjunction with high-throughput sequencing, for rapid development of vaccines against previously unknown pathogens. The results of high-throughput scans may also have relevance to whole-genome scans of potential causes of autoimmune dis-

eases.

Peptide-based vaccine design and cancer immunotherapy offer a number of interesting algorithmic challenges besides the problems addressed in this work. First, we focus on a single step, MHC binding, in the antigen processing pathway only. More sophisticated approaches address the other relevant steps as well, i.e. proteasomal cleavage or TAP transport. Second, the identification of adequate epitopes is only the initial step in the design of a peptide vaccine. Due to the high variability in the MHC genes, good epitopes will vary from patient to patient. Furthermore, the number of different peptides one can reasonably include in a vaccine is very limited (see e.g. the discussion in the review [7]). Hence, an economically interesting peptide vaccine should contain a minimal number of epitopes covering the majority of the alleles encountered in the whole population. One way to reduce the number of peptides required to cover most MHC alleles is the use of promiscuous epitopes, i.e. peptides which are epitopes for multiple alleles.

The identification of such an optimal set can be formulated as an optimization problem where the “average immunogenicity” of the vaccine is maximized. This average immunogenicity can be computed from the distribution of the alleles for a given population and the relative immunogenicity (estimated through the binding strength) of each peptide for a specific allele. In the context of personalized medicine, a variant of this optimization problem is to identify the optimal set of peptides for a patient’s immunotype.

Even more complex optimization problems arise through the possible optimization of the proposed peptides (either in the anchor residues or in the region recognized by the T-cell receptor). Due to the combinatorial complexity of this problem, it can be addressed using sophisticated optimization techniques only. It thus represents another exciting challenge for future work.

Acknowledgments: We thank Kari Irvine and Ruobing Wang for their helpful feedback on the biology of T-cell epitopes and the needs of the laboratory with regards to prediction methods. We also thank Von Bing Yap for his work on determining important properties predictive of binding affinity. O.K. is currently at Universität des Saarlandes. R.S. is currently at Carnegie Mellon University. S.H. is currently at Sanaria.

References

- [1] A. Bairoch and R. Apweiler. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucl. Acids Res.*, 28:45–48, 2000.
- [2] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and B. A. Rapp. GenBank. *Nucl. Acids Res.*, 28(1):15–18, 2000.
- [3] C. Branden and J. Tooze. *Introduction to Protein Structure*. Garland Publishing, Inc, New York, 1991.
- [4] V. Brusic, G. Rudy, and L. C. Harrison. MHCPEP, a database of MHC-binding peptides. *Nucl. Acids Res.*, 26(1):368–371, 1998.
- [5] C. Burge and S. Karlin. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, 268:78–94, 1997.
- [6] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–167, 1998.
- [7] C. Buteau, S. Markovic, and E. Celis. Challenges in the development of effective vaccines for cancer. *Mayo Clin Proc*, 77:339–349, 2002.
- [8] P. Dönnes and A. Elofsson. Prediction of MHC class I binding peptides, using SVMHC. *BMC Bioinformatics*, 3:25, 2002.
- [9] M. S. Gelfand, E. V. Koonin, and A. A. Mironov. Prediction of transcription regulatory sites in Archaea by a comparative genomic approach. *Nucl. Acids Res.*, 28:695–705, 2001.
- [10] C. Gianfrani, C. Oseroff, J. Sidney, R.W. Chesnut, and A. Sette. Human memory ctl response specific for influenza a virus is broad and multi-specific. *Human Immun.*, 61:438–452, 2000.
- [11] M. Gribskov, A. D. McLachlan, and D. Eisenberg. Profile analysis: detection of distantly related proteins. *Proc. Natl. Acad. Sci. USA*, 84:4355–4358, 1987.
- [12] A. S. De Groot, A. Bosma, N. Chinai, J. Frost, B. M. Jesdale, M. A. Gonzalez, W. Martin, and C. Saint-Aubin. From genome to vaccine: in silico predictions, ex vivo verification. *Vaccine*, 19:4385–4395, 2001a.
- [13] A. S. De Groot, C. Saint-Aubin, A. Bosma, J. Rayner, and W. Martin. Rapid determination of HLA B*07 ligands from the West Nile virus NY99 genome. *Emerging Infections Diseases*, 7(4):706–713, 2001b.

- [14] K. Gulukota, J. Sidney, A. Sette, and C. DeLisi. Two complementary methods for predicting peptides binding major histocompatibility complex molecules. *J. Mol. Biol.*, 267:1258–1267, 1997.
- [15] D. F. Hunt, R. A. Henderson, J. Shabanowitz, K. Sakaguchi, H. Michel, N. Sevilir, et al. Three-dimensional structure of a peptide extending from one end of a class I MHC binding site. *Science*, 255:1261–63, 1992.
- [16] S. Kawashima, H. Ogata, and M. Kanehisa. AAindex: amino acid index database. *Nucleic Acids Res.*, 27:368–369, 1999.
- [17] S.L. Lauemoller, C. Kesmir, S.L. Corbet, A. Formsgaard, A. Holm, M.H. Claesson, S. Brunak, and S. Buus. Identifying cytotoxic t cell epitopes from genomic and proteomic information: "The human MHC project". *Rev Immunogenet.*, 2:477–91, 2000.
- [18] H. Mamitsuka. Predicting peptides that bind to MHC molecules using supervised learning of hidden markov models. *Proteins: Struct. Func. Gen.*, 33(4):460–74, 1998.
- [19] Matlab. The Mathworks, Natick Massachusetts.
- [20] K. C. Parker, M. A. Bednarek, and J. E. Coligan. Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. *J. Immunol.*, 152(10):163–175, 1994a.
- [21] H. Rammensee, J. Bachmann, N.P. Emmerich, O.A. Bachor, and S. Stevanovic. SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogen.*, 50(3-4):213–9, 1999.
- [22] N. Renkvist, C. Castelli, P. F. Robbins, and G. Parmiani. A listing of human tumor antigens recognized by T cells. *Cancer Immunol. Immunother.*, 50:3–15, 2001.
- [23] D. Rognan, S. L. Lauemoller, A. Holm, S. Buus, and V. Tschinke. Predicting binding affinities of protein ligands from three-dimensional models: application to peptide binding to class I major histocompatibility proteins. *J. Med. Chem.*, 42:4650–4658, 1999.
- [24] S. A. Rosenberg. A new era for cancer immunotherapy based on the genes that encode cancer antigens. *Immunity*, 10:281–287, 1999.
- [25] J. Ruppert, J. Sidney, E. Celis, R.T. Kubo, H.M. Grey, and A. Sette. Prominent role of secondary anchor residues in peptide binding to HLA-A2.1 molecules. *Cell*, 74:929–937, 1993.
- [26] J. A. Schafer, B. M. Jesdale, J. A. George, N. M. Koutatab, and A.S. De Groot. Prediction of well-conserved HIV-1 ligands using a matrix-based algorithm, EpiMatrix. *Vaccine*, 16(19):1880–4, 1998.
- [27] T. D. Schneider, G. D. Stormo, L. Gold, and A. Ehrenfraucht. Information content of binding sites on nucleotide sequences. *J. Mol. Biol.*, 188:415–431, 1986.
- [28] M. Singh and P. Kim. Towards predicting coiled-coil protein interactions. In *Proceedings of the 3rd Annual International Conference on Computational Molecular Biology (RECOMB-01)*, pages 279–286, Montreal, 2001. ACM press.
- [29] G. D. Stormo. DNA binding sites: representation and discovery. *Bioinformatics*, 16:16–23, 2000.
- [30] G. D. Stormo and G. W. Hartzell. Identifying protein-binding sites from unaligned DNA fragments. *Proc. Natl. Acad. Sci. USA*, 86:1183–1187, 1989.
- [31] J. D. Thompson, D. G. Higgins, and T. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties, and weight-matrix choice. *Nucl. Acids Res.*, 22:4673–4680, 1994.
- [32] J.C. Venter, M.D. Adams, E.W. Myers, P.W. Li, R.J. Mural, G.G. Sutton, et al. The sequence of the human genome. *Science*, 291:1304–1351, 2001.
- [33] C. H. Wu, H. Huang, L. Arminski, J. Castro-Alvear, Y. Chen, Z.-Z. Hu, et al. The Protein Information Resource: an integrated public resource of functional annotation of proteins. *Nucl. Acids Res.*, 30(1):35–37, 2002.