BIOINFORMATICS

Vol. 00 no. 00 2006
Pages 1–7

Identifying HLA Supertypes by Learning Distance Functions

Tomer Hertz [a,b], Chen Yanover [a]

[a]School of Computer Science and Engineering, The Hebrew University of Jerusalem, Israel,
[b]Interdisciplinary Center for Neural Computation, The Hebrew University of Jerusalem,Israel

ABSTRACT

The development of epitope-based vaccines crucially relies on the ability to classify Human Leukocyte Antigen (HLA) molecules into sets that have similar peptide binding specificities, termed *supertypes*. In their seminal work, Sette and Sidney [21] defined 9 HLA class I supertypes, and claimed that these provide an almost perfect coverage of the entire repertoire of HLA class I molecules.

HLA alleles are highly polymorphic and polygenic and therefore experimentally classifying each of these molecules to supertypes is at present an impossible task. Recently, a number of computational methods have been proposed for this task. These methods are based on defining protein similarity measures, derived from analysis of binding peptides or from analysis of the proteins themselves [13, 7]. In this paper we define both peptide derived and protein derived similarity measures, which are based on learning distance functions. The peptide driven measure is defined using a peptide-peptide distance function, which is learnt using information about known binding and non-binding peptides [28]. The protein similarity measure is defined using a protein-protein distance function, which is learnt using information about alleles previously classified into supertypes by [21]. We compare the classification obtained by these two complimentary methods to previously suggested classification methods. In general, our results are in excellent agreement with the classifications suggested by Sette and Sidney [21] and with those reported in [13].

There are two important advantages of our proposed distance-based approach. First, it makes use of two different and important immunological sources of information – HLA alleles and peptides that are known to bind or not bind to these alleles. Second, since each of our distance measures is trained using a different source of information, their combination can provide a more confident classification of alleles into supertypes.

1 INTRODUCTION

The major task of recognizing foreign pathogen proteins is mediated by interactions between Major Histocompatibility Complex (MHC) molecules and short pathogen-derived peptides. When such a peptide binds to an MHC molecule, the complex is transported to the cell surface, where it can be recognized by T-cells that in turn elicit an immune response. Predicting protein-peptide binding is therefore of central importance for developing peptide-based (or epitope-based) vaccines. These vaccines contain only selected subsequences, or epitopes, derived from an entire protein which are known to bind to various MHC molecules. There are several important advantages of epitope-based vaccines: they can induce more potent immune responses and they appear to be safer to use and easier to produce [21, 20, 22, 13].

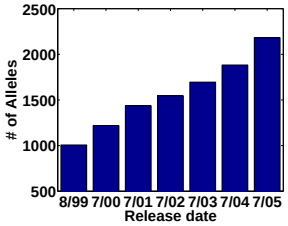

Fig. 1. The growth rate of the number of alleles in the IMGT/HLA database [18].

Designing epitope-based vaccines with high population coverage is a challenging problem for the following two main reasons. First, MHC molecules are highly selective and only bind to specific peptides – each molecule binds to approximately 1% of all existing peptides [29]. Their binding specificity is determined by the molecular structure and the chemical properties of the MHC binding sites. Second, MHC molecules are highly polymorphic and polygenic. Currently, the IMGT/HLA database [18] (version 2.9), lists 1245 HLA class I and 744 HLA class II alleles are known (1046 HLA class I and 604 HLA class II proteins). Each individual carries only a few alleles (up to 6 HLA class I alleles and up to 12 HLA class II alleles) [12]. One of the ways to overcome this large degree of polymorphism is to make use of epitopes which bind to a number of HLA molecules.

Luckily, it turns out that despite the high polymorphism exhibited by HLA alleles, many HLA molecules bind to sets of overlapping peptides. These alleles can be grouped into *supertypes* — sets of alleles that bind to similar peptides. Identifying these supertypes is therefore an important task with clear implications to the development of epitope-based vaccines. However, experimentally determining binding specificity of a **single** allele is a hard task which requires both rigorous experimental validation and theoretical analysis. Tackling this problem for more than 1600 proteins is currently impractical [7]. Additionally, as may be seen in Fig. 1, around 200 new HLA alleles are discovered each year. The difficulty of the task at hand and the rapid increase in the number of alleles call for the development of computational tools for HLA supertype classification.

Recently, a number of computational methods have been proposed for HLA supertype classification [17, 13, 7]. These methods first define a protein-protein similarity (or equivalently a distance) measure. This similarity measure is then

used to classify the proteins into supertypes. Although the supertype classification problem is defined over proteins, its underlying goal of finding peptides which bind to several proteins, requires exploration of the "peptide space". This inherent duality leads to two different approaches for defining the similarity between proteins:

"Peptide-based" approaches define a similarity measure between proteins that is based on properties of sets of **peptides** that bind to these proteins. The binding peptides used can either be experimentally determined binders [21] or computationally predicted binders (e.g based on binding motifs) [17]. The similarity between these sets of binding peptides can be defined by counting the overlapping peptides [21, 17] or by representing each set by a motif, and then defining some similarity measure over these motifs [13].

"Protein-based" approaches define a similarity measure that is based on the properties of the proteins themselves. It has been noted that proteins that bind to a set of overlapping peptides have binding sites that are similar to one another [7]. By using some canonical representation of the binding sites, one can define a similarity measure over these binding sites.

In this paper we present a novel approach for computationally identifying supertypes, that is based on learning distance functions. Unlike previous works, we address the supertype classification problem using the two complementary approaches described above — both peptide and protein-based similarities. More specifically, we propose to explicitly learn distance functions of two different types:

A peptide-peptide distance function using information about binding and non-binding peptides. We have recently presented a framework for protein-peptide binding prediction, based on learning a peptide-peptide distance function over an entire family of proteins (e.g. HLA class I) [28]. In our current work, we show how this distance function can be naturally used to define a distance measure over proteins.

A protein-protein distance function using information about alleles which have been experimentally determined to belong to the same supertype (e.g. by [21]). We propose to use the same distance learning algorithm to directly learn protein-protein distance functions.

Using these two complementary methods we classify a set of HLA-A and HLA-B alleles into supertypes. We then compare the results obtained using these two methods to previously proposed methods and characterize their regions of agreement/disagreement. We believe that comparing these two methods, each trained using a different source of information, may shed new light on the supertype classification problem.

1.1 Related work

HLA class I supertypes were originally defined by Sette, Sidney and colleagues during the second half of the 1990's [5, 24, 23, 21]. In these seminal works, supertype classification was essentially based on overlapping sets of peptides, known to bind to a subset of HLA alleles. Additionally, the classification took into account properties of the binding pockets of these alleles. All in all, 9 HLA supertypes were identified [21]: $A1$, $A2$, $A3$ and $A24$ for HLA-A alleles; $B7$, $B27$, $B44$, $B58$ and $B62$ for HLA-B alleles.

Recently, several computational methods for defining supertypes have been proposed. Lund et. al [13] construct Hidden Markov Models (HMMs) for HLA class I molecules using a Gibbs sampling procedure. They then define a similarity measure between these sequence motifs and use this similarity to cluster alleles into supertypes. Reche and Reinherz [17] rank a set of 1000 random peptides, using a position specific scoring matrix for each protein, and then consider the top 2% scoring peptides as predicted binders. They define a similarity between alleles that is based on counting the number of peptides that are predicted to bind to both alleles. As in [13], they then cluster the alleles using the neighbor clustering algorithm from the Phylogeny Inference Package (PHYLIP) [8]. Both [13] and [17] define supertypes, based on similarity between sets of binding peptides, either directly [17] or using some representation for each set of peptides (e.g. sequence motifs).

A somewhat complimentary approach is taken by Doytchinova et. al [7], who classify supertypes by examining the binding sites of various HLA-A, HLA-B and HLA-C alleles. Each allele was modeled based on a reference x-ray structure and its binding site residues used to define a set of characteristic properties. They then use both a hierarchical clustering algorithm and Principal Component Analysis (PCA) to cluster these alleles into supertypes.

It is important to note that despite the fact that the supertype problem is of central importance, to date there is no clear ground truth classification of alleles into supertypes. In all of the works described above, and in our current work, the results are compared to the supertypes defined by Sette and Sidney [21]. Their classification only includes a small subset of currently known HLA class I alleles, and therefore most alleles are currently unlabeled.

2 LEARNING PEPTIDE DISTANCE FUNCTIONS FOR SUPERTYPE CLASSIFICATION

For sake of completeness, we first present our novel framework for protein-peptide binding prediction based on learning peptide-peptide distance functions. We then show how these learnt distance functions can be naturally used to define a protein-protein similarity measure for supertype classification.

2.1 Learning peptide-peptide distance functions

Previously proposed learning approaches for protein-peptide binding prediction, address the binding prediction problem using traditional margin based binary classifiers: for each protein a classifier is trained to distinguish binding peptides from non-binding peptides [6, 4, 16] (for a review see [9]). Recently, we proposed *PepDist*: a novel approach for predicting binding affinity based on learning peptide-peptide distance functions[1] [28]. Our approach is based on two important observations:

[1] A distance function is a function that assigns some non-negative value for each pair of points (or peptides in our case).

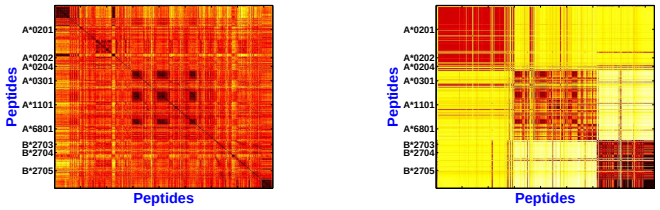

Fig. 2. Peptide-peptide distance matrices of HLA binding peptides, collected from the MHCPEP dataset [3]. The data consists of peptides known to bind to 9 different HLA class I proteins (see labels on the y-axis). Peptides that bind to each of the proteins were grouped together and labeled accordingly. In both matrices the value in position (i, j) represents the distance between $Peptide_i$ and $Peptide_j$ (the darker the color is, the smaller the distance). A "good" peptide-peptide distance matrix should therefore be block diagonal. **(a)** Naive peptide-peptide distance matrix (Euclidean distance in $\mathbb{R}^{45}$). **(b)** The peptide-peptide distance matrix learnt using the *DistBoost* algorithm. *DistBoost* was trained on binding peptides from all of the proteins **simultaneously**.

OBSERVATION 1. *Peptides that bind to the same protein are similar to one another, and different from non-binding peptides.*

OBSERVATION 2. *Peptides binding to different proteins within the same "family" resemble each other*

Observation 1 implicitly underlies most, if not all, computational prediction methods. A direct implication of this observation is that the distance between a query peptide and a set of known binders can be used to predict the peptide's binding affinity: A peptide that is close to the set of binders would be classified as a binder and a peptide far from this set would be classified as a non-binder. We therefore suggested to predict protein-peptides binding affinity by learning a distance function over pairs of peptides [28]. Moreover, based on *observation 2*, we suggested to learn a **single** peptide-peptide distance function over an **entire** family of proteins (e.g MHC class I). This distance function can be used to compute the affinity of a novel peptide to any of the proteins in the given family. Note that the mere definition of supertypes is based on the latter observation. The learnt peptide-peptide distances can therefore also be used to identify supertypes by clustering together proteins that bind to similar peptides.

Recently, there has been a growing interest in the problem of learning distance functions in the machine learning community. Most algorithms that learn distance functions make use of equivalence constraints [10, 11, 1, 27, 2, 26]. Equivalence constraints are relations between pairs of data points, which indicate whether the points in the pair belong to the same category or not. We term a constraint *positive* when the points are known to be from the same class, and *negative* in the opposite case. In this setting, the goal of the algorithm is to learn a distance function that complies with the equivalence constraints provided as input. Specifically, in our previous work [28] and in this work, we use the *DistBoost* algorithm which is a semi-supervised distance learning algorithm [10, 11]. *DistBoost* learns a distance function using a well known machine learning technique, called *Boosting* [19]. For details regarding the algorithm's description see [10, 28].

We formalize the peptide-peptide distance learning problem as follows: each protein is denoted by some class label. Each pair of peptides, which are known to bind to a specific protein (that is, belong to the same class), defines a positive constraint, while each pair of peptides in which one binds to the protein and the other does not — defines a negative constraint. Therefore, for each protein, our training data consists of a list of binding and non-binding peptides, and a set of equivalence constraints that they induce. We collect these sets of peptides and equivalence constraints for several proteins within a protein family into a **single** dataset. We then use this dataset to learn a peptide-peptide distance function (see Fig. 2 (b)). Using this distance function, we can predict the binding affinity of a novel peptide to a specific protein, by measuring its average distance to all of the peptides which are known to bind to that protein (see Fig. 3 (b)). We have tested our approach on binding prediction of MHC class I and MHC class II datasets and in all cases our method provided excellent results which outperformed most state-of-the-art computational prediction methods [28].

2.2 Using peptide distance functions to define supertypes

As noted above, the definition of supertypes is based on identifying HLA molecules that bind to sets of overlapping peptides. This definition implies that the distances between binding peptides can be naturally used to define a distance measure over pairs of proteins. This protein-protein distance function can then be used to classify proteins into supertypes. A "good" distance function would assign small distances to proteins which bind to sets of overlapping peptides, and large distances to proteins which bind to different (or non-overlapping) sets of peptides. One intuitive way to formalize this notion is to use the average distance between peptides that bind to two different proteins as a measure of their similarity. More formally, let us denote by $D_{peptides}(Peptide_i, Peptide_j)$ the distance between $Peptide_i$ and $Peptide_j$. We define the distance between $Protein_m$ and $Protein_n$, $D_{proteins}(Protein_m, Protein_n)$ to be:

$$D_{proteins}(Protein_m, Protein_n) \equiv \frac{1}{N_{mn}} \sum_{i \in B_n, j \in B_m} D_{peptides}(Peptide_i, Peptide_j) \quad (1)$$

where B_n and B_m denote the sets of peptides known to bind to $Protein_n$ and $Protein_m$ respectively and $N_{nm} = |B_n| \cdot |B_m|$.

Figure 3 (c) presents an illustrative example of the protein-protein distance matrix between 9 different HLA class I alleles using the peptide-peptide distance matrix in Figure 3 (a) (see also previous section). As can be clearly seen, distances between proteins from the same supertype are smaller than the distances between proteins from different supertypes. In order to classify alleles into supertypes we can now

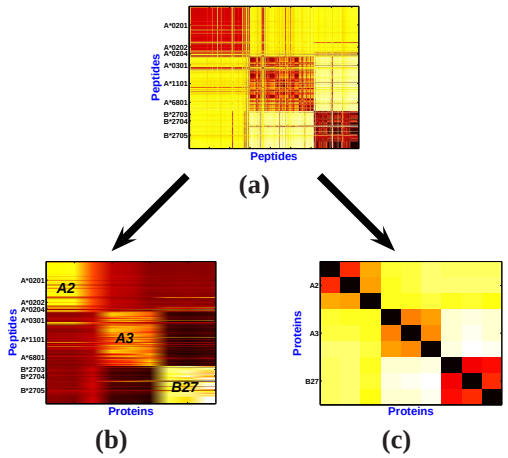

Fig. 3. The Peptide-peptide distance function **(a)** can be used to provide both protein-specific binding prediction (b) and to identify supertypes (c). **(b)**: a protein-peptide affinity matrix where the value in position (i, j) is the predicted binding affinity of $Peptide_i$ to $Protein_j$. Peptides that bind to each of the 9 proteins were grouped together. The first three proteins belong to the $A2$ supertype, the next three proteins to the $A3$ supertype and the last three to the $B27$ supertype. As may be seen the affinity values of peptides that are known to bind to a specific protein are higher than their affinity to other proteins. Also note that the binding affinities of peptides to different proteins within the same supertype are very similar. **(c)**: The protein-protein distance matrix defined using peptide distances (eq. 1). Proteins are ordered as described above. As can be seen, the distances between proteins within each of the three supertypes are smaller than the distances between proteins from different supertypes.

feed this protein distance matrix into any generic hierarchical clustering algorithm (e.g. average linkage) and identify clusters of alleles as supertypes. We used this peptide-based similarity measure to classify a set of HLA-A and HLA-B proteins into supertypes and present the results obtained in Section 4.

3 LEARNING A DISTANCE FUNCTION OVER PROTEIN BINDING SITES FOR DEFINING SUPERTYPES

In the previous section we defined supertypes based on peptide-peptide distance functions, learned using information about binding and non-binding peptides for a family of proteins. An alternative, somewhat complementary approach, is to directly define a similarity measure between the proteins themselves. As noted by [7], the binding sites of two proteins that bind to similar peptides, share common characteristics. Such similarities were also analyzed in the original works of Sette and Sidney (see e.g. [21]), who used this information to suggest tentative assignment of alleles to various HLA class I supertypes.

Our distance learning algorithm can also be used to learn a distance function over the binding sites of the various HLA alleles. In order to learn such a distance function we need to define a representation of the binding sites for each allele and to show how to make use of experimentally validated supertype classification to define equivalence constraints. Following Doytchinova et. al [7], we defined the binding sites of HLA-A alleles using a set of 35 amino acids and HLA-B binding sites as a set of 37 amino acids. We extract equivalence constraints as follows: each pair of proteins which are known to belong to the same supertype (based on the classification of [21]), forms a positive constraint and each pair of proteins which are known to belong to different supertypes forms a negative constraint. Note that we only have information regarding a small subset of all known alleles, and therefore (as in the peptide-peptide distance learning scenario) the distance learning scenario is semi-supervised.

4 EXPERIMENTS AND RESULTS

We now present the results obtained by our peptide- and protein-based approaches, both using the same distance learning framework, on classification of HLA class I alleles into supertypes. We begin with a short description of our experimental setup including the datasets we used, data representation and algorithmic details.

4.1 Experimental setup

Datasets Sequences of 9 amino acid long (9-mers) peptides, that are known to bind the HLA class I proteins listed in [13], were collected from the MHCPEP [3] and SYFPEITHI [15] datasets. Peptides, that contain undetermined residues (denoted by the letter code X), were excluded. We then grouped all 9-mers, that bind to HLA class I molecules (both HLA-A and HLA-B), to a **single** dataset, called *HLA1peptides*. The *HLA1peptides* dataset includes 4273 peptides that bind to 112 HLA class I alleles (42 HLA-A alleles and 70 HLA-B alleles).

Sequences of HLA alleles were acquired from the IMGT/HLA Sequence Database ([18], version 2.9). In this version, 372 HLA-A and 661 HLA-B class I alleles have been named. We obtained a set of unique proteins, considering only the first allele (denoted by $*xxxx01$) out of each group of amino acid identical alleles [7]. We then defined a dataset of HLA-A alleles, called *HLA-Aproteins*, and a dataset of HLA-B alleles, called *HLA-Bproteins*. The former dataset consists of 301 HLA-A alleles and the latter - 573 HLA-B proteins. Following [7], the HLA-A binding site includes 35 residues and the HLA-B binding site consisted of 37 residues. These binding site definitions are based on x-ray structures of reference proteins. HLA alleles were then aligned within each locus, using the initial x-ray structure as a template.

To allow a fair comparison between all methods, we used the list of alleles presented by Lund et. al. [13] and compared the classifications reported by all methods on this subset only. We therefore present a comparison of all methods on a subset of 95 HLA class I alleles. In all comparisons to [7] we used the results reported using a hierarchical clustering algorithm.

Data representation *DistBoost*, the distance learning algorithm used in this paper, requires that the data be represented in some continuous vector feature space. As in our previous work [28] we used the representation suggested by [25]. Using Venkatarajan and Braun's feature vectors, we represent each 9 amino acid long peptide as a point in $\mathbb{R}^{45}$, by simply concatenating its amino-acid feature vectors. Similarly, we encode each sequence of residues defining HLA-A

and HLA-B allele binding site as a 175 and 185 dimensional vector respectively. The vectors representing the binding sites, were further processed using PCA to obtain vectors in $\mathbb{R}^{50}$.

Extracting equivalence constraints The distance learning algorithm we used is trained using unlabeled data and additional equivalence constraints. Extracting equivalence constraints for our peptide-based approach is rather straightforward: each pair of peptides that are known to bind to the same protein form a positive equivalence constraint. Extracting equivalence constraints for our proposed protein-based approach was done as described in Sec. 3: each pair of proteins which were grouped into the same supertype by Sette and Sidney [21], formed a positive equivalence constraint and each pair of proteins which are known to belong to different supertypes forms a negative constraint. It is important to note that a rather small portion of the binding-site datasets were tagged using these constraints – 30 out of the 301 in the HLA-Aproteins dataset and 85 out of the 573 in the HLA-Bproteins dataset. Only 20 HLA-A and 31 HLA-B of these tagged alleles appear in the set of 95 alleles on which we provide a detailed comparison with previously suggested methods.

Algorithmic setup The *DistBoost* algorithm was run for 100 iterations on the peptide dataset and 30 iterations on the binding site datasets. In order to cluster the alleles into supertypes we used two standard clustering methods: (1) The average-linkage algorithm and (2) The *Neighbor* program from the PHYLIP package [8] (as in [13, 17]). Each cluster was automatically assigned a label as follows: We begin by labelling each point using the classification in [13] (of which the labels of [21] are almost in full agreement). We then labeled each cluster using the most frequent label within the cluster. The number of clusters chosen was equivalent to the number of supertypes described in [13]. In order to verify that our labeling scheme was sensible, we also drew dendrograms of our peptide- and protein-based methods and visually inspected them after coloring each node with our predicted classifications. Dendrograms were drawn using the TreeView program [14].

4.2 Supertype classification results

Table 2 presents a comparison of the classification of alleles to supertypes obtained by our two proposed methods and the classifications reported by [21, 13, 7]. Each supertype was colored using a different color. The percentage of agreement between the various methods are summarized in Table 1. In these calculations we ignore entries in which one of the two methods did not provide a classification (entries marked with $'?'$ or with '—')[2]. As can be seen, our methods are in high agreement with all other compared methods. The agreement between our two methods is 75%. We can also see that our protein-based prediction is in 94% agreement with the classifications of Sette and Sidney [21]. This is not very surprising,

[2] When comparing our results to those suggested by Sette and Sidney, we do not consider alleles labeled as $A26$ and $B39$ as disagreeing with the labels $A1$ and $B27$ respectively. The $A26$ and $B39$ supertypes are two novel supertypes that were recently suggested by Lund et. al. [13].

Methods	% Agreement
[21] Vs. [13]	99 %
BS. Dist Vs. [13]	97 %
BS. Dist Vs. [21]	94 %
Pep. Dist Vs. [21]	82 %
Pep. Dist Vs. [13]	77 %
Pep. Dist Vs. BS. Dist	75 %
[21] Vs. [7]	49 %
[13] Vs. [7]	46 %
Pep. Dist Vs. [7]	44 %
BS. Dist Vs. [7]	41 %

Table 1. Comparative supertype classification results. The bars present the percentage of agreement between different methods. As may be seen, our two methods are in high agreement with the classifications provided by [21, 13].

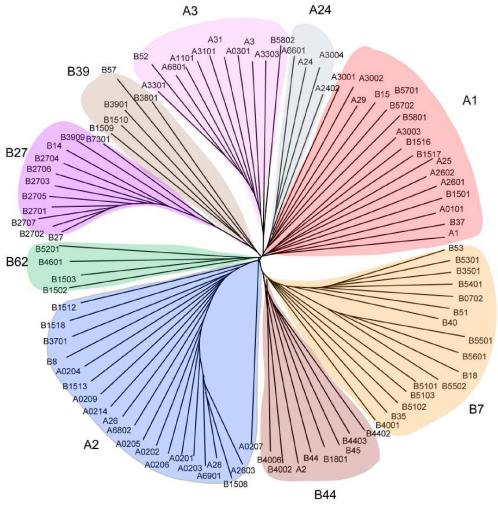

Fig. 4. A dendrogram plot showing the classification obtained using our peptide-based approach. Each supertype is marked with a different color and therefore results are best seen in color.

due to the fact that when training our algorithm on the binding site datasets, we used the classifications provided by Sette and Sidney as constraints provided to our distance learning algorithm. It is interesting to note that our peptide-based method which was not provided with any constraints regarding the proteins, still obtains an 82% agreement with the results of Sette and Sidney.

In order to further visualize the results obtained by our peptide-based approach, Fig. 4 presents a colored dendrogram of the classifications obtained by our method. Each cluster is associated with a specific supertype. Interestingly, we can see that most of the alleles on which our method disagrees with the classifications of other methods are near cluster boundaries (see for example $B4001$ which our peptide-based approach classified as belonging to the $B7$ supertype, and is classified by most other methods to the adjacent $B44$ supertype). Additionally, many of these involve serologically defined specificities (e.g. $A2$, $B57$ etc.), as opposed to genetically defined specificities. Serologically defined species, are usually based on earlier experiments, and their motifs are not always definitive. For example, the $A28$ allele consists of two different specificities as a results of a crossover event, and was therefore later divided into $A68$

Allele	Pept. Dist.	BS Dist.	[21] Class.	[13] Class.	[7] Class.	Allele	Pept. Dist.	BS Dist.	[21] Class.	[13] Class.	[7] Class.	Allele	Pept. Dist.	BS Dist.	[21] Class.	[13] Class.	[7] Class.
A*0101	A1	A1	A1	A1	A3	A*6801	A3	A3	A3	A3	A3	B*3801	B39	B39	B27	B39	B44
A*0201	A2	A2	A2	A2	A2	A*6802	A2	A2	A2	A2	A2	B*3901	B39	B39	B27	B39	B7
A*0202	A2	A2	A2	A2	A2	A*6901	A2	A2	A2	A2	A2	B*3909	B27	B39	?	B39	B7
A*0203	A2	A2	A2	A2	A2	B*0702	B7	B7	B7	B7	B7	B40	B7	—	?	B44	—
A*0204	A2	A2	A2	A2	A2	B08	A2	—	?	?	—	B*4001	B7	B44	B44	B44	B27
A*0205	A2	A2	A2	A2	A2	B14	B27	—	?	?	—	B*4002	B44	B44	B44	B44	B27
A*0206	A2	A2	A2	A2	A2	B15	A1	—	?	B62	—	B*4006	B44	B44	B44	B44	B27
A*0207	A2	A2	A2	A2	A2	B*1501	A1	B62	B62	B62	B27	B44	B44	—	B44	B44	—
A*0209	A2	A2	A2	A2	A2	B*1502	B62	B62	B62	B62	B7	B*4402	B44	B44	B44	B44	B44
A*0214	A2	A2	A2	A2	A2	B*1503	B62	B27	B27	B27	B27	B*4403	B44	B44	B44	B44	B44
A3	A3	—	A3	A3	—	B*1508	B7	B7	B7	B7	B7	B45	B44	—	?	?	—
A*0301	A3	A1	A3	A3	A3	B*1509	B39	B39	B27	B39	B7	B*4601	B62	B62	B62	B62	B27
A*1101	A3	A1	A3	A3	A3	B*1510	B39	B39	B27	B39	B7	B51	B7	—	B7	B7	—
A24	A24	—	A24	A24	—	B*1512	A2	B62	B62	B62	B27	B*5101	B7	B7	B7	B7	B44
A*2402	A24	A24	A24	A24	—	B*1513	A2	B62	B62	B62	B44	B*5102	B7	B7	B7	B7	B44
A25	A1	—	?	A1	—	B*1516	A1	B58	B58	B58	B44	B*5103	B7	B7	B7	?	B44
A26	A2	—	?	A26	—	B*1517	A1	B58	B58	B58	B44	B52	A3	—	?	B62	—
A*2601	A1	A26	A1	A26	A2	B*1518	A2	B27	B27	B27	B7	B*5201	B62	B62	?	B62	B44
A*2602	A1	A26	A1	A26	A2	B18	B7	—	?	B44	—	B53	B7	—	?	B7	—
A*2603	A2	A26	?	A26	A2	B*1801	B44	B7	B44	?	B7	B*5301	B7	B7	B7	B7	B44
A28	A2	—	?	A26	—	B27	B27	—	B27	B27	—	B*5401	B7	B7	B7	B7	B7
A29	A1	—	?	?	—	B*2701	B27	B27	B27	B27	B44	B*5501	B7	B7	B7	B7	B7
A*2902	B44	A1	?	?	A3	B*2702	B27	B27	B27	B27	B44	B*5502	B7	B7	B7	B7	B7
A*3001	A1	A1	A24	A1	A3	B*2703	B27	B27	B27	B27	B27	B*5601	B7	B7	B7	B7	B7
A*3002	A1	A1	?	A1	A3	B*2704	B27	B27	B27	B27	B27	B57	B58	—	?	B58	—
A*3003	A1	A1	?	A1	A3	B*2705	B27	—	B27	B27	B27	B*5701	A1	B58	B58	B58	B44
A*3004	A24	A1	?	A1	A3	B*2706	B27	B27	B27	B27	B27	B*5702	A1	B58	B58	B58	B44
A31	A3	—	?	?	—	B*2707	B27	B27	B27	B27	B27	B*5801	A24	B58	B58	B58	B44
A*3101	A3	—	A3	A3	A3	B35	B7	—	?	B7	—	B*5802	B58	B58	B58	B58	B44
A*3301	A3	A3	A3	A3	A3	B*3501	B7	B7	B7	B7	B7	B*7301	B27	B27	B27	B27	B7
A*3303	A3	A3	A3	A3	A3	B37	A1	—	?	?	—						

Table 2. Comparative supertype classification results. In general, the results obtained by our proposed peptide-based (Pept. Dist) and protein-based (BS Dist.) approaches are in agreement with the classifications suggested by [21, 13, 7]. Each supertype is colored using a different color and therefore results are best seen in color. '?' denotes unlabeled alleles. '—' marks alleles which did not appear in the *HLA-Aproteins* and *HLA-Bproteins* datasets.

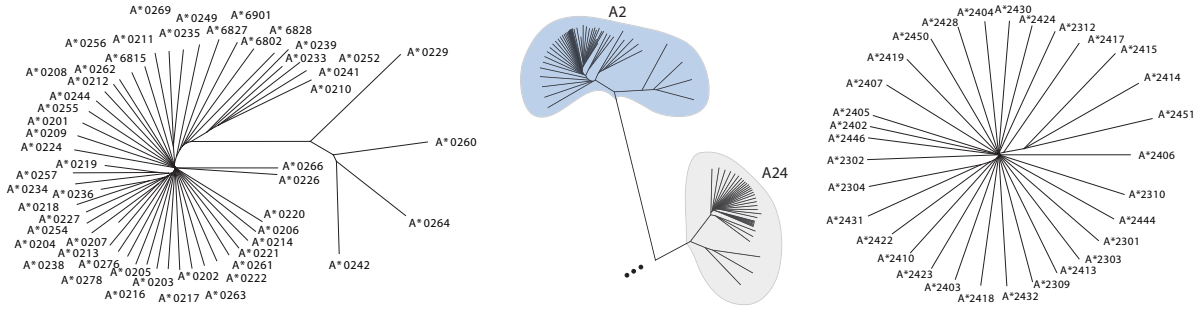

Fig. 5. **(b)** Part of the dendrogram plot showing the classification of 301 HLA-A alleles obtained by our protein-based approach. **(a)** A blowup of the A2 cluster in (b). **(c)** A blowup of the A24 cluster in (b).

and $A69$ (J. Sidney, personal communication). Since in our current paper we wanted to provide a clear comparison to previously suggested methods, we evaluated our method on the exact same 95 alleles that were used in [13]. In future work, it may therefore be beneficial to exclude the serological alleles from supertype classification studies.

A partial visualization of the results obtained by our protein-based method is shown in Fig. 5. Since the datasets used to train the binding site distance functions contained a large number of alleles, it is very hard to visualize the clustering of the entire dataset. Two distinct supertypes are shown: $A2$ and the $A24$. To the best of our knowledge, many of the alleles in these clusters have not been previously classified. Examples are the $A0261$, $A0262$, $A0263$, $A6827$ and $A6828$ alleles which our method classifies to the $A2$ supertype and the $A2444$ and $A2312$ alleles which are predicted to belong to the $A24$ supertype.

5 DISCUSSION AND CONCLUDING REMARKS

In this paper, we presented a novel framework for classification of HLA alleles into supertypes which is based on learning distance functions. We showed how the same algorithmic solution can be utilized for learning both peptide and protein distance functions and hence provide peptide- and protein-based approached for supertype classification. Our novel method can make use of the two most important sources of immunological information which are continuously collected and published: the HLA alleles themselves and the peptides that are known to bind (and also not-bind) to these alleles.

Since most known alleles are not clearly classified into supertypes, it is quite difficult to provide a quantitative measure of performance. Comparing our results to previously suggested methods, showed that our binding site prediction method was in excellent agreement with most other methods,

and our peptide-based approach was in good agreement with most other methods (including the binding-site method). Our protein-based approach also provided supertype classification predictions for new, previously unclassified alleles. Despite the fact that our peptide-based approach provided only an 82% agreement with the results of Sette and Sidney, it is important to note that it did not make use of any information regarding the classification of alleles into supertypes. This result clearly conforms with the original observations that led to the definition of supertypes, as alleles that bind to overlapping sets of peptides. Our binding site results show, that it is also feasible to classify alleles into supertypes by directly measuring the similarity between their binding sites.

The idea of using both a protein-based classification and a peptide-based classification method proposes the following combined supertype classification and binding prediction process: when a novel allele is discovered (for which no binding peptides are currently known), a protein-based approach can be used to classify it into one of the currently identified supertypes. This classification can now help guide experiments which seek to identify peptides that bind to this novel protein. These binding peptides, can in turn be used to retrain a peptide-based classifier and to obtain an additional classification of the new allele to a supertype. Since the two methods rely on different sources of information, we should expect more confident classifications when their predictions agree with one another. Therefore, our proposed novel distance-based approach not only allows a more confident classification of alleles into supertypes, but can also guide experimental HLA-peptide binding assays.

In our future work, we hope to incorporate additional data for training our peptide-based method. Of special interest is the incorporation of experimentally determined non-binders, which were currently not used. In our previous work [28], we have shown that learning a peptide distance function using additional information about non-binders, provides better prediction results. We therefore hope that using this additional source of information will further improve our classification results.

Acknowledgements

We thank John Sidney for many useful comments and suggestions. C.Y is supported by Yeshaya Horowitz Association through the Center for Complexity Science.

REFERENCES

[1]A. Bar-Hilel, T. Hertz, N. Shental, and D. Weinshall. Learning distance functions using equivalence relations. In *The 20th International Conference on Machine Learning.*, 2003.

[2]M. Bilenko, S. Basu, and R.J Mooney. Integrating constraints and metric learning in semi-supervised clustering. In *International Conference on Machine Learning, ICML 2004*, Banff Canada, 2004. AAAI press.

[3]V Brusic, G Rudy, and L. C. Harrison. MHCPEP, a database of MHC-binding peptides: update 1997. *Nucl. Acids Res.*, 26(1):368–371, 1998.

[4]S. Buus, S.L. Lauemoller, P. Worning, C. Kesmir, T. Frimurer, S. Corbet, A. Fomsgaard, J. Hilden, A. Holm, and S. Brunak. Sensitive quantitative predictions of peptide-MHC binding by a 'query by committee' artificial neural network approach. *Tissue Antigens*, 62(5):378–384, 2003.

[5]MF del Guercio, J Sidney, G Hermanson, C Perez, HM Grey, RT Kubo, and A Sette. Binding of a peptide antigen to multiple HLA alleles allows definition of an A2-like supertype. *J Immunol*, 154(2):685–693, 1995.

[6]P. Donnes and A. Elofsson. Prediction of MHC class I binding. *BMC Bioinformatics*, 3, 2002.

[7]I. A. Doytchinova, P Guan, and D. R. J Flower. Identifiying human MHC supertypes using bioinformatic methods. *Journal of Immunology*, 172:4314–4323, 2004.

[8]J Felsenstein. PHYLIP (phylogeny inference package) version 3.5c. Distributed by the author, 1993.

[9]D. R. Flower. Towards in silico prediction of immunogenic epitopes. *TRENDS in immunology*, 24, 2003.

[10]T. Hertz, A. Bar-Hillel, and D. Weinshall. Boosting margin based distance functions for clustering. In *ICML*, 2004.

[11]T. Hertz, A. Bar-Hillel, and D. Weinshall. Learning distance functions for image retrieval. In *CVPR, Washington DC, June 2004*, 2004.

[12]Charles A. Janeway, Paul Travers, Mark Walport, and Mark Shlomchik. *Immunobiology*. New York and London: Garland Publishing, 5th edition, 2001.

[13]Ole Lund, Morten Nielsen, Can Kesmir, AndersGorm Petersen, Claus Lundegaard, Peder Worning, Christina Sylvester-Hvid, Kasper Lamberth, Gustav Rder, Sune Justesen, Sren Buus, and Sren Brunak. Definition of supertypes for HLA molecules using clustering of specificity matrices. *Immunogenetics*, 55(12):797–810, 2003-4.

[14]R. D. M. Page. Treeview: An application to display phylogenetic trees on personal computers. *Computer Applications in the Biosciences*, 12:357–358, 1996.

[15]H. G. Rammensee, J. Bachmann, N. P. N. Emmerich, O. A. Bachor, and S. Stevanovic. SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics*, 50(3-4):213–219, 1999.

[16]P. A. Reche, J. P. Glutting, H Zhang, and E. L. Reinherz. Enhancement to the RANKPEP resource for the prediction of peptide binding to MHC molecules using profiles. *Immunogenetics*, 26(6):405–419, 2004.

[17]P. A. Reche and E. L. Reinherz. Definition of MHC supertypes through clustering of MHC peptide binding repertoires. In *Lecture Notes in Computer Science ICARIS 2004*, volume 3239, pages 189–196, 2004.

[18]James Robinson, Matthew J. Waller, Peter Parham, Natasja de Groot, Ronald Bontrop, Lorna J. Kennedy, Peter Stoehr, and Steven G. E. Marsh. IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex. *Nucl. Acids Res.*, 31(1):311–314, 2003.

[19]Robert E. Schapire and Yoram Singer. Improved boosting using confidence-rated predictions. *Machine Learning*, 37(3):297–336, 1999.

[20]A. Sette, M. Newman, B. Livingston, D. McKinney, J. Sidney, G. Ishioka, S. Tangri, J. Alexander, J. Fikes, and R. Chesnut. Optimizing vaccine design for cellular processing, MHC binding and TCR recognition. *Tissue Antigens*, 59(6):443–443, 2002.

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

[22]Alessandro Sette and John Fikes. Epitope-based vaccines: an update on epitope identification, vaccine design and delivery. *Current Opinion in Immunology*, 15(4):461–470, 2003.

[23]J Sidney, S Southwood, MF del Guercio, HM Grey, RW Chesnut, RT Kubo, and A Sette. Specificity and degeneracy in peptide binding to HLA-B7-like class I molecules. *J Immunol*, 157(8):3480–3490, 1996.

[24]John Sidney, Howard M. Grey, Scott Southwood, Esteban Celis, Peggy A. Wentworth, Marie-France del Guercio, Ralph T. Kubo, Robert W. Chesnut, and Alessandro Sette. Definition of an HLA-A3-like supermotif demonstrates the overlapping peptide-binding repertoires of common HLA molecules. *Human Immunology*, 45(2):79–93, 1996.

[25]M. S. Venkatarajan and W. Braun. New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical-chemical properties. *Journal of Molecular Modeling*, 7(12):445–453, 2001.

[26]Jean-Philippe Vert and Yoshihiro Yamanishi. Supervised graph inference. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 1433–1440. MIT Press, Cambridge, MA, 2005.

[27]E.P Xing, A.Y. Ng, M.I. Jordan, and S. Russell. Distance metric learnign with application to clustering with side-information. In *Advances in Neural Information Processing Systems*, volume 15. The MIT Press, 2002.

[28]Chen Yanover and Tomer Hertz. Predicting protein-peptide binding affinity by learning peptide-peptide distance functions. In *The Ninth Annual Conference on Research in Computational Biology - RECOMB 2005*, 2005.

[29]Jonathan W. Yewdell and Jack R. Bennink. Immunodominance in Major Histocompatibility Complex Class I-Restricted T-Lymphocyte Responses. *Annual Review of Immunology*, 17:51–88, 1999.