A novel representation of HLA allele specificity and prediction algorithm for HLA class I binding.

David K. Crockett1,2, Stephen R. Piccolo2, Perry G. Ridge1, Hernando Escobar1, Eduardo Reyes-Vargas1, Peter E. Jensen1, Julio C. Delgado1
1ARUP Laboratories, Department of Pathology, University of Utah School of Medicine, Salt Lake City, UT 2Department of Biomedical Informatics, University of Utah School of Medicine, Salt Lake City, UT 84109

INTRODUCTION

Current software tools and algorithms for prediction of HLA class I peptide binding prediction are typically based on the frequency of 2 anchoring residues from peptides of 8 to 10 residues long.

For many years now, this representation of HLA class I peptide motifs described by dominant amino acid residues located in primary anchor positions has been widely used. For example, the reported motif for HLA-A*0201 from the SYFPEITHI is "x-[LM]-x-x-x-x-x-x-[VL]." (Rammensee, Friede et al. 1995).

Variations of this representation are also seen in many other HLA class I peptide motif databases such as IMGT/HLA (Robinson, Malik et al. 2000) and IEDB (Peters, Sidney et al. 2005).

This has led to the development of software tools and algorithms for peptide binding prediction and screening of target organisms or sequences for a given peptide motif.

Although this current motif representation allows for relative frequency of anchoring residues to be detailed, less account is given for variable peptide length (8-mer, 9-mer, 10-mer, etc.) and properties of amino acid residues internal to anchor positions.

Based on high confidence mass spectrometric sequencing of eluted HLA peptides, we herein propose a novel representation that not only models HLA motifs based on amino acid properties of key peptide anchor position residues but also physicochemical properties of molecular spacing of internal residues between anchor positions.

Figure 1 displays this novel representation of HLA allele specificity. Representative peptides (n=773) from 8 HLA class I subtypes were identified by LC-MS/MS using the LC-MS/MS method from Escobar et al. (Escobar, Crockett et al. 2008). A publicly available data set of A*0201 binding peptides (n=1181) and non-binding peptides (n=1908) was also downloaded from IEDB. (Peters, Sidney et al. 2005)

Unique A*0201 peptides were characterized for motif anchor residues and anchor spacing using values from the 544 physical, chemical, conformational, or energetic properties found in AAIndex v9.4 (http://www.genome.jp/aaindex/). (Kawashima and Kanehisa 2000)

The most relevant properties were identified using Correlation-based Feature Subset Selection algorithm (Hall 2000), together with the Best First (greedy hill-climbing) search method.

Several classification models were then trained and tested for their ability to discriminate HLA A*0201 binders from non-binders for the IEDB and MS/MS confirmed peptides.

Algorithms evaluated for predicting HLA allele specificity included:

- IBk (k nearest neighbor)
- JRip (rules)
- NaiveBayes (Bayesian)
- RandomForest (trees)
- SimpleLogistic (regression)
- SMO (support vector machine)
- ZeroR (zero rules)

Attribute selection and classification algorithms were implemented using the Weka software package v3.6. (Frank 2005)

Previously reported measurements of algorithm performance including sensitivity, specificity and positive predictive value were calculated. (Hand 2009)

RESULTS

For HLA-A*0201, supervised feature selection of amino acid properties in AAIndex found 12 best descriptors for Anchor 1 and 7 for Anchor 2. Internal spacing between the anchors was best characterized by 6 attributes.

Classification performance (positive predictive value) for A*0201 ranged from 61% to 81% for the seven classifier algorithms evaluated.

The best performing classification (Figure 2) identified A*0201 binders (n=629) with >94% accuracy.

This novel representation of HLA class I motifs yields unique biochemical descriptors of binding preference in class I peptides – independent of the length of the peptide.

CONCLUSIONS

We have developed a novel algorithm for improved representation of HLA Class I motifs.

This model was used to accurately classify A*0201 from 8 other HLA supertypes.

This novel representation of HLA allele specificity models a unique and more detailed biochemical description for properties of binding preference in MHC class I peptide anchor residues and the internal molecular spacing between those anchors.

Mass spectrometric confirmation of HLA class I peptide motifs is a key tool in generating data sets for advancing HLA prediction algorithms.

Future work will include allele representation and binding prediction tools for additional MHC class I subtypes.

References: