Development and validation of a broad scheme for prediction of HLA class II restricted T cell epitopes

Sinu Paul[1], John Sidney[2], Bjoern Peters[3], Alessandro Sette[4]
La Jolla Institute for Allergy & Immunology
9420 Athena Cir,
La Jolla, CA 92037 USA
001 (858) 752-6916
1.spaul@liai.org, 2.jsidney@liai.org, 3.bpeters@liai.org, 4.alex@liai.org

ABSTRACT

Computational prediction of HLA class II restricted T cell epitopes has great significance in many immunological studies including vaccine discovery. With the development of novel bioinformatics approaches, prediction of HLA class II binding has improved significantly but a strategy to predict the most dominant HLA class II epitopes has not been defined. Using different sets of peptides from various allergen and bacterial antigens and HLA class II binding prediction tools from the IEDB, we have designed a strategy to predict the top epitopes from any antigen. We found that the top 21% of 15-mer peptides overlapping by 10 residues (based on the predicted binding to seven DRB1 and DRB3/4/5 alleles) capture 50% of the immune response. This corresponded to an IEDB consensus percentile rank of 19.82 which could be used as a universal prediction threshold.

Categories and subject descriptors

J.3 [Life and Medical Sciences]: biology and genetics, health.

General Terms

Algorithms, Measurement, Documentation, Design.

Keywords

Epitope prediction, HLA class II restricted T cell epitopes

1. INTRODUCTION

Prediction and identification of HLA class II restricted T cell epitopes is a task of significance for several applications, including studying the immune response against pathogens or allergens, or deimmunization of protein-based drugs. Class II molecules are alpha/beta heterodimers encoded by four different loci in humans: DRB1/DRA, DRB1/DRB3/4/5, DPA/DPB and DQA/DQB. HLA class II polymorphism presents a challenge for epitope identification [9]. However, the problem is simplified by focusing on the alleles most frequently expressed in the general population, which account for an overwhelming majority of the expressed genes [4]. Furthermore, extensive similarities exist within the peptides bound by different allelic variants [3]. Accordingly, peptides capable of binding multiple HLA class II molecules (i.e., promiscuous peptides) often account for a large fraction of antigen specific T cell responses [6, 8]. While computational predictions of HLA class II binding capacity have improved as novel bioinformatic approaches have been implemented [5, 8, 12], a strategy to predict the most dominant HLA class II epitopes has not been defined.

2. MATERIALS AND METHODS

2.1 Immunogenicity studies

Sets of overlapping 15 or 16mer peptides spanning various allergen and bacterial antigens were screened for immune reactivity as previously described [1, 6, 7] (Supplemental Table 1). Antigen specific cytokine production induced in donor PBMC was measured in dual ELISPOT assays. Responses to timothy grass, cockroach and house dust mite peptides were measured following in vitro stimulation with respective allergen extracts, and responses to *Bordetella pertussis* antigens following stimulation with corresponding peptide pools. Responses to mycobacterial antigens were analyzed ex vivo. Peptide specific responses were expressed as spot-forming cells (SFCs)/10^6 peripheral-blood mononuclear cells (PBMCs). Donors were HLA typed at each class II locus to four-digit resolution by SSP or deep sequencing methods [11].

2.2 Prediction of binding affinity

Peptide affinity for HLA class II alleles was predicted using the MHC II binding prediction tool available at the IEDB (www.iedb.org) [10, 11]. Allele-specific consensus percentile ranks of all algorithms queried by the IEDB tool were utilized [12]. A percentile rank is generated by comparing the selected peptide's predicted binding affinity against that of a large set of similarly sized peptides randomly selected from the SWISS-PROT database. Percentile rank provides a uniform scale allowing comparisons across different predictors. A lower percentile rank indicates higher affinity. In the case of consensus method, median of the percentile ranks of the three methods involved is considered as the IEDB consensus percentile rank.

2.3 Correction for epitope redundancy

Responses against two consecutive peptides are often due to the same minimal epitope. To avoid counting the same epitope twice, two consecutive responses with magnitudes within 2.5-fold of each other were merged into a single antigenic region, and the higher SFC value utilized. The region is considered successfully

BCB '14, September 20 - 23 2014, Newport Beach, CA, USA

ACM 978-1-4503-2894-4/14/09$15.00.
http://dx.doi.org/10.1145/2649387.2660842

predicted if either of the two peptides is predicted, and “credit” for prediction is given only once.

3. RESULTS AND DISCUSSION

3.1 Definition of alternative prediction strategies for HLA class II epitopes

Predicted binding of peptides in the data sets (Supplemental Table 1) for a previously described set of 26 HLA class II alleles that are most frequent in the general worldwide population [3] (Supplemental Table 2), was determined as described above. To evaluate the efficacy of various approaches to identify the most dominant epitope responses, employing these predictions, the percentage (fraction) of peptides in each data set needed to capture 50% of the total response (50% of the total SFC values in the data set - expressed as SFCs/10^6 PBMCs) was utilized as a performance metric.

As a first approach, we considered the “promiscuous binding capacity” of each peptide which is determined by the number of alleles bound (i.e., peptides binding with more alleles being better binders; and peptides with predicted IEDB consensus percentile rank ≤20 are called as “binding” to that allele). Using this approach, as shown in Figure 1a-b, an average of 30.91% (range 25.35%-40.08%) peptides were needed to capture 50% of the total response in the data set. In the second approach, we considered the “median consensus percentile rank” of each peptide, which is defined as the median of the IEDB consensus percentile ranks predicted for the set of 26 selected alleles. This approach was the most effective, with the top 26.26% (range 16.90%-38.38%) of the peptides capturing 50% of the total response (Figure 1a, c).

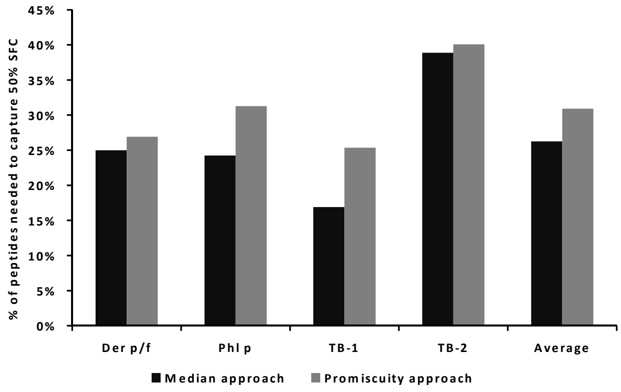

Figure 1a: Performance of two approaches for implementing HLA class II binding predictions to identify T cell epitopes. The % of peptides needed for each method to identify a panel of epitopes accounting for 50% of the total antigen specific response (SFC) is shown for 4 different systems, as described in the text: Der p/f (House dust mite), Phl p (Timothy grass), TB-1 (*Mycobacterium tuberculosis*), and TB-2 (*M. tuberculosis*). Black bars show performance based on ranking peptides according to the median consensus percentile rank against a panel of the 26 most common HLA class II alleles. The grey bars show performance based on ranking peptides according to the number of alleles predicted to bind (promiscuity). A lower % of peptides indicates better performance.

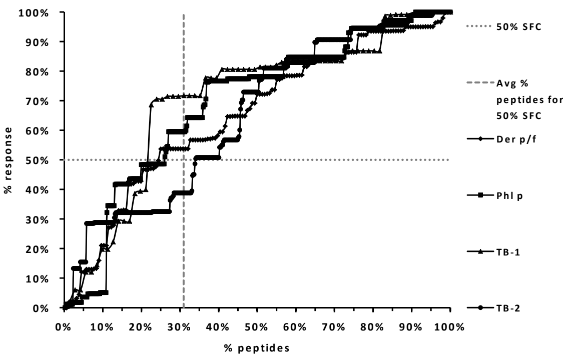

Figure 1b: The % of response as a function of % of peptides using the “promiscuity” approach for the four data sets. An average of 30.91% peptides was needed to capture 50% SFC with this approach.

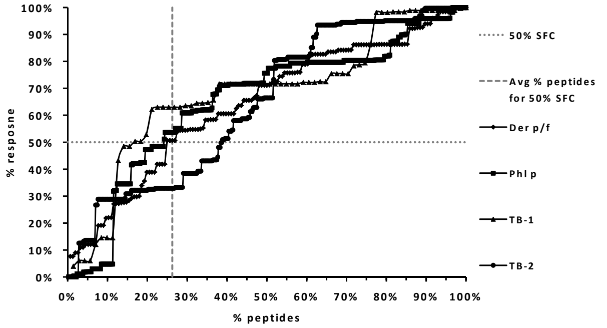

Figure 1c: The % of response as a function of % of peptides using the “median consensus percentile rank” approach for the four data sets. An average of 26.26% peptides was needed to capture 50% SFC with this approach.

3.2 Exclusion of DP locus improves predictive efficacy

As different HLA class II loci appear to contribute differentially to human responses [6], we hypothesized that examining the performance as a function of the class II locus may improve predictions. The average % of peptides required to capture 50% SFC for different combinations of DRB1, DRB3/4/5, DQ, and DP alleles are shown in Figure 2. The best results (23.82%) were obtained when DP alleles were left out. The lower performance of methods incorporating DP molecules might be due to the fact that less binding data is available for these molecules leading to worse prediction algorithms.

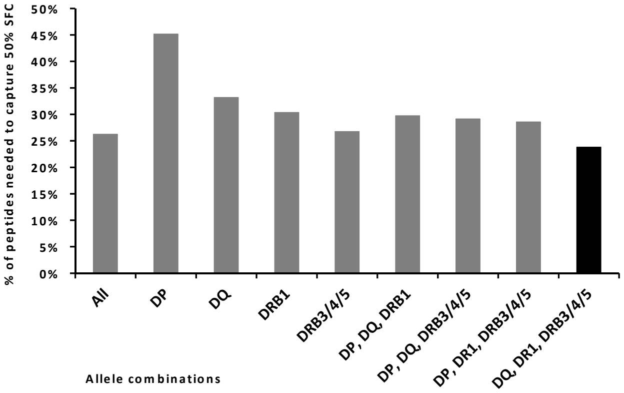

Figure 2: Performance of the "median consensus percentile rank" approach as a function of variable inclusion of the DP, DQ, DRB1 and DRB3/4/5 loci.

3.3 Optimal results obtained with a set of seven DRB1 and DRB3/4/5 alleles

We next examined the effect of varying the specific alleles included in the prediction panel. Frequency thresholds for inclusion were varied independently for each locus (i.e., DQ, DRB1 and DRB3/4/5). The best results (21.41% of peptides needed to capture 50% SFC) were observed when the three DRB1 alleles with frequency ≥ 12% (DRB1*03:01, DRB1*07:01, DRB1*15:01) were used along with the four DRB3/4/5 alleles (DRB3*01:01, DRB3*02:02, DRB4*01:01, DRB5*01:01) (data not shown). This empirical optimization is probably reflective of the fact that DR alleles are the most dominant locus restricting HLA class II responses in humans. It is noteworthy that the seven allelic variants cover the main HLA class II supertypes [3].

3.4 Predictions based on alleles frequent in specific donor cohorts

The HLA composition of a donor varies with their ethnicity. Accordingly, we examined the performance of the "median consensus percentile rank" predictions utilizing alleles present with frequencies of 10% or more in each specific donor population. As above, the best results (19.69%) were obtained using only DRB1 and DRB3/4/5 alleles (Figure 3). At the same time, the improvement seen here is minor suggesting that tailoring the prediction to a specific population has limited value.

3.5 Defining a universal prediction threshold

The percent of total peptides required to capture 50% of the response, as calculated here on a protein-by-protein basis is not available when considering individual peptides. To derive a standard prediction threshold, we calculated the median IEDB consensus percentile rank, using predictions for the seven DRB1 and DRB3/4/5 alleles highlighted above, associated with the selected set of peptides yielding 50% of the response. This value was found to be 19.82 (median consensus percentile rank from the seven selected alleles).

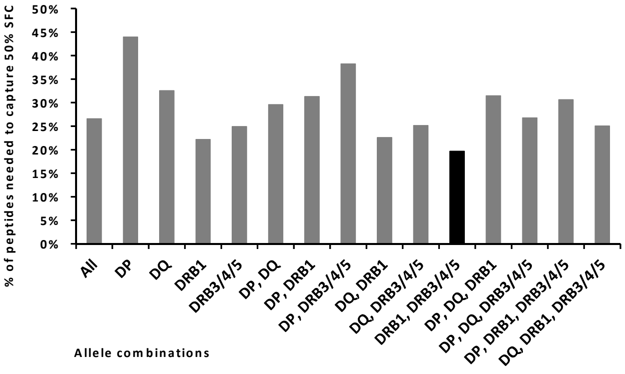

Figure 3: Average % of peptides required to capture 50% SFC for different allele combinations using the alleles with frequency >10% in each specific corresponding donor cohort.

3.6 Validation of the results with blind prediction using new data sets

The analyses above suggested that the optimal approach for efficient selection of epitope candidates would be based on determining the median consensus percentile rank across a selected panel of seven DR alleles (3 DRB1 alleles with frequency ≥12% in conjunction with 4 DRB3/4/5 alleles). To validate these results we examined overlapping peptides for two additional sets of proteins of immunological interest: 1) cockroach allergens and 2) pertussis vaccine antigens. When the range of approaches tried above was implemented against the two blind sets, the best performance was again achieved with the "median consensus percentile rank" approach. When the universal median IEDB consensus percentile threshold defined above (~20%) with the panel of seven DR alleles was utilized, the average % of SFC captured using peptides with median IEDB consensus percentile rank ≤ 20.0 was found to be 48.55%, confirming the validity of this prediction threshold.

4. CONCLUSIONS

We scrutinized the use of HLA class II binding predictions to identify sets of epitopes with high immunological activity. The results validate previous observations that promiscuous binders account for a high fraction of the total response. Compared to HLA class I predictions, the results are sobering, as this performance is remarkably less effective. This is in line with other recent studies [2]. At the same time, our results provide guidance for practical implementation of predictions, and identify specific subsets of HLA molecules that are most effectively considered by prediction schemes. The synthesis of approximately 20% of the peptides in a set of 15-mer peptides overlapping by 10 residues allows covering a 200-residue protein (otherwise covered by 40 overlapping peptides) with 10 peptides, which still constitutes significant cost savings.

5. ACKNOWLEDGEMENTS

We thank Cecilia Arlehamn, Myles Dillon, Daniela Weiskopf, Denise Hinz, Carla Oseroff and Veronique Schulten for help with the immunogenicity screening and analyses and Denise McKinney for the HLA typing analyses. This project has been funded with federal funds from the NIAID under Contract No.

HHSN272201200010C (IEDB), HHSN272200900044C (TB), and HHSN272200700048C (Allergy), in addition to Grant No. NIH U19 AI100275 (Allergy).

6. REFERENCES

[1] Arlehamn, C. S., Sidney, J., Henderson, R., Greenbaum, J. A., James, E. A., Moutaftsi, M., Coler, R., McKinney, D. M., Park, D., Taplitz, R., Kwok, W. W., Grey, H., Peters, B. and Sette, A. 2012. Dissecting mechanisms of immunodominance to the common tuberculosis antigens ESAT-6, CFP10, Rv2031c (hspX), Rv2654c (TB7.7), and Rv1038c (EsxJ*). J Immunol*, 188, 10, 5020-5031.

[2] Chaves, F.A., Lee, A.H., Nayak, J.L., Richards, K.A. and Sant, A.J. 2012. The Utility and Limitations of Current Web-Available Algorithms To Predict Peptides Recognized by CD4 T Cells in Response to Pathogen Infection. *J Immunol*, 188, 9, 4235-4248.

[3] Greenbaum, J., Sidney, J., Chung, J., Brander, C., Peters, B. and Sette, A. 2011. Functional classification of class II human leukocyte antigen (HLA) molecules reveals seven different supertypes and a surprising degree of repertoire sharing across supertypes. *Immunogenetics*, 63, 325-335.

[4] McKinney, D. M., Southwood, S., Hinz, D., Oseroff, C., Arlehamn, C. S., Schulten, V., Taplitz, R., Broide, D., Hanekom, W. A., Scriba, T. J., Wood, R., Alam, R., Peters, B., Sidney, J. and Sette, A. 2013. A strategy to determine HLA class II restriction broadly covering the DR, DP, and DQ allelic variants most commonly expressed in the general population. *Immunogenetics*, 65, 5, 357-370.

[5] Nielsen, M., Lund, O., Buus, S. and Lundegaard, C. 2010. MHC Class II epitope predictive algorithms. *Immunology*. 130(3), 319-328.

[6] Oseroff, C., Sidney, J., Kotturi, M. F., Kolla, R., Alam, R., Broide, D. H., Wasserman, S. I., Weiskopf, D., McKinney, D. M., Chung, J. L., Petersen, A., Grey, H., Peters, B. and Sette, A. 2010. Molecular determinants of T cell epitope recognition to the common Timothy grass allergen. *J Immunol*, 185, 2, 943-955.

[7] Oseroff, C., Sidney, J., Tripple, V., Grey, H., Wood, R., Broide, D. H., Greenbaum, J., Kolla, R., Peters, B., Pomes, A. and Sette, A. 2012. Analysis of T cell responses to the major allergens from German cockroach: epitope specificity and relationship to IgE production. *J Immunol*, 189, 2, 679-688.

[8] Paul, S., Kolla, R. V., Sidney, J., Weiskopf, D., Fleri, W., Kim, Y., Peters, B. and Sette, A. 2013. Evaluating the immunogenicity of protein drugs by applying in vitro MHC binding data and the immune epitope database and analysis resource. *Clin Dev Immunol*, 2013, 467852.

[9] Robinson, J., Waller, M. J., Parham, P., de Groot, N., Bontrop, R., Kennedy, L. J., Stoehr, P. and Marsh, S. G. 2003. IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex. *Nucleic Acids Res*, 31, 1, 311-314.

[10] Vita, R., Zarebski, L., Greenbaum, J. A., Emami, H., Hoof, I., Salimi, N., Damle, R., Sette, A. and Peters, B. 2010. The immune epitope database 2.0. *Nucleic Acids Res*, 38, Database issue, D854-862.

[11] Wang, C., Krishnakumar, S., Wilhelmy, J., Babrzadeh, F., Stepanyan, L., Su, L. F., Levinson, D., Fernandez-Vina, M. A., Davis, R. W., Davis, M. M. and Mindrinos, M. 2012. High-throughput, high-fidelity HLA genotyping with deep sequencing. *Proc Natl Acad Sci USA*, 109, 22, 8676-8681.

[12] Wang, P., Sidney, J., Kim, Y., Sette, A., Lund, O., Nielsen, M. and Peters, B. 2010. Peptide binding predictions for HLA DR, DP and DQ molecules. *BMC Bioinformatics*, 11, 568.

[13] Zhang, Q., Wang, P., Kim, Y., Haste-Andersen, P., Beaver, J., Bourne, P. E., Bui, H. H., Buus, S., Frankild, S., Greenbaum, J., Lund, O., Lundegaard, C., Nielsen, M., Ponomarenko, J., Sette, A., Zhu, Z. and Peters, B. 2008. Immune epitope database analysis resource (IEDB-AR). *Nucleic Acids Res*, 36, Web Server issue, W513-518.

Supplemental Table 1: Data sets used in the analyses

Data set	No. of antigens	Antigens	Peptides	Total peptides	No. of donors	Ethnicity	Reference
Der p/f	4	proDer p 1.0105	59	156*	20	White	Oseroff, in preparation
		proDer f 1.0101	59				
		Der p 2.0101	24				
		Der f 2.0103	24				
Phl p	10	Phl p 1	51	425	25	Mixed (predominantly white)	Oseroff, 2010
		Phl p 2	23				
		Phl p 3	18				
		Phl p 4	103				
		Phl p 5.0103	61				
		Phl p 6	26				
		Phl p 7	14				
		Phl p 11	27				
		Phl p 12	25				
		Phl p 13	77				
TB-1	4	ESAT-6	17	71	18	Mixed	Arlehamn, 2012
		CFP10	18				
		hspX	27				
		Rv1038c	9				
TB-2	11	Rv0125	69	499	32	Coloured	Arlehamn, in preparation McKinney, 2012
		Rv0288	18				
		Rv1196	77				
		Rv1813c	27				
		Rv1886c	63				
		Rv2608	114				
		Rv2660c	13				
		Rv3619	17				
		Rv3620c	18				
		Rv3804c	66				
		Rv3875	17				
Cockroach	6	Bla g 1	189	463	19	Mixed (predominantly black)	Oseroff, 2012
		Bla g 2	69				
		Bla g 4	35				
		Bla g 5	39				
		Bla g 6	76				
		Bla g 7	55				
Pertussis	9	fhaB	468	785	23	Mixed	Dillon, in preparation
		fim2	26				
		fim3	25				
		prn	131				
		ptxA	40				
		ptxB	30				
		ptxC	28				
		ptxD	21				
		ptxE	16				

*10 peptides are shared by different antigens in the Der p/f data set. Thus, the total no. of unique peptides is 10 less than the sum of peptides in individual antigens.

Supplemental Table 2: The 26 MHC class II alleles that are most frequent in the general worldwide population [3], and thus were included in the analyses.

Locus	Allele	Phenotype frequency	Gene frequency
DRB1	DRB1*0101	5.4	2.8
	DRB1*0301	13.7	7.1
	DRB1*0401	4.6	2.3
	DRB1*0405	6.2	3.1
	DRB1*0701	13.5	7.0
	DRB1*0802	4.9	2.5
	DRB1*0901	6.2	3.1
	DRB1*1101	11.8	6.1
	DRB1*1201	3.9	2.0
	DRB1*1302	7.7	3.9
	DRB1*1501	12.2	6.3
	Total	71.1	46.2
DRB3/4/5	DRB3*0101	26.1	14.0
	DRB3*0202	34.3	18.9
	DRB4*0101	41.8	23.7
	DRB5*0101	16.0	8.3
	Total	87.7	65.0
DQA1/DQB1	DQA1*0501/DQB1*0201	11.3	5.8
	DQA1*0501/DQB1*0301	35.1	19.5
	DQA1*0301/DQB1*0302	19.0	10.0
	DQA1*0401/DQB1*0402	12.8	6.6
	DQA1*0101/DQB1*0501	14.6	7.6
	DQA1*0102/DQB1*0602	14.6	7.6
	Total	81.6	57.1
DPB1	DPB1*0101	16.0	8.4
	DPB1*0201	17.5	9.2
	DPB1*0401	36.2	20.1
	DPB1*0402	41.6	23.6
	DPB1*0501	21.7	11.5
	Total	92.6	72.7