Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.

Journal of Immunological Methods 374 (2011) 53–61

Contents lists available at ScienceDirect

Journal of Immunological Methods

journal homepage: www.elsevier.com/locate/jim

Research paper

MULTIPRED2: A computational system for large-scale identification of peptides predicted to bind to HLA supertypes and alleles

Guang Lan Zhang [a,*], David S. DeLuca [a], Derin B. Keskin [a], Lou Chitkushev [b], Tanya Zlateva [b], Ole Lund [c], Ellis L. Reinherz [a], Vladimir Brusic [a]

[a] Cancer Vaccine Center, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA
[b] Department of Computer Science, Metropolitan College, Boston University, Boston MA, USA
[c] Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Lyngby, Denmark

ARTICLE INFO

Article history:
Received 3 September 2010
Received in revised form 29 October 2010
Accepted 18 November 2010
Available online 2 December 2010

Keywords:
T-cell epitope hotspots
HLA
HLA supertype
Human leukocyte antigen
Promiscuous binding peptide
Vaccine design

ABSTRACT

MULTIPRED2 is a computational system for facile prediction of peptide binding to multiple alleles belonging to human leukocyte antigen (HLA) class I and class II DR molecules. It enables prediction of peptide binding to products of individual HLA alleles, combination of alleles, or HLA supertypes. NetMHCpan and NetMHCIIpan are used as prediction engines. The 13 HLA Class I supertypes are A1, A2, A3, A24, B7, B8, B27, B44, B58, B62, C1, and C4. The 13 HLA Class II DR supertypes are DR1, DR3, DR4, DR6, DR7, DR8, DR9, DR11, DR12, DR13, DR14, DR15, and DR16. In total, MULTIPRED2 enables prediction of peptide binding to 1077 variants representing 26 HLA supertypes. MULTIPRED2 has visualization modules for mapping promiscuous T-cell epitopes as well as those regions of high target concentration – referred to as T-cell epitope hotspots. Novel graphic representations are employed to display the predicted binding peptides and immunological hotspots in an intuitive manner and also to provide a global view of results as heat maps. Another function of MULTIPRED2, which has direct relevance to vaccine design, is the calculation of population coverage. Currently it calculates population coverage in five major groups in North America. MULTIPRED2 is an important tool to complement wet-lab experimental methods for identification of T-cell epitopes. It is available at http://cvc.dfci.harvard.edu/multipred2/.

© 2010 Elsevier B.V. All rights reserved.

1. Introduction

T cells identify foreign antigens through their T-cell receptor (TCR), which interacts with a peptide antigen in complex with a major histocompatibility complex (MHC) molecule in conjunction with CD4 or CD8 co-receptors (Meuer et al., 1982; Wang and Reinherz, 2002). For example, CD8$^+$ T cells control viral infection through direct cytolysis of infected cells and through production of soluble antiviral mediators. This function is mediated by linear peptide epitopes presented by MHC class I molecules. CD4$^+$ T cells recognize epitopes presented by MHC class II molecules on the surface of virus-infected cells and secrete lymphokines that stimulate B cells and cytotoxic T cells. The recognition of a given antigenic peptide by the immune system of an individual depends on the peptide's ability to bind one or more of the host's human leukocyte antigens (HLA, human MHC). There is a great diversity of HLA genes with more than 5000 known variants characterized as of April 2010 (Robinson et al., 2009). HLA proteins share three-dimensional structures with main differences observed in residues that form the peptide binding groove (Bjorkman et al., 1987). HLA proteins that have small differences in their peptide binding grooves and share similar peptide binding specificities are grouped into HLA supertypes (Sette and Sidney, 1999; Lund et al., 2004). Promiscuous peptides, i.e., those that bind multiple HLA variants, are suitable targets for peptide-based vaccine development because they are relevant for diverse HLA populations. Immunological hotspots, defined as regions comprising clusters of promiscuous T-cell epitopes, have been

[*] Corresponding author. Tel.: +1 582 7951; fax: +1 617 632 3351.
E-mail address: guanglan_zhang@dfci.harvard.edu (G.L. Zhang).

0022-1759/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jim.2010.11.009

determined in some antigens, such as SARS coronavirus nucleocapsid (Gupta et al., 2006), HIV-1 proteins (Surman et al., 2001; Brown et al., 2003), or *Chlamydia trachomatis* outer membrane protein (Kim and DeMars, 2001). These clusters are suitable vaccine targets for the development of epitope-based vaccines. Such vaccines focus on a small number of selected hotspots that can potentially elicit required T-cell activation through multiple HLA molecules. Wet-lab experiments are time-consuming and costly and their applicability for large-scale screening is limited. Computational tools are essential for identifying T-cell epitopes and immunological hotspots for development of population-based vaccines. They are normally used for pre-screening of targets, followed by experimental validation using small number of well-selected target peptides.

Several online computational systems were previously developed to address various issues related to selection of potential promiscuous T-cell epitopes. MULTIPRED is a computational system for prediction of promiscuous HLA binding peptides to HLA-A2, -A3, -B7, and -DR supertypes (Zhang et al., 2005; Zhang et al., 2007). PEPVAC (Promiscuous EPitope-based VACcine) is a web server for multi-epitope vaccine development based on the prediction of supertypic MHC ligands (Reche and Reinherz, 2005). It predicts promiscuous peptide binders to five HLA class I supertypes, A2, A3, B7, A24, and B15. It also estimates the phenotypic population frequency of these supertypes. Hotspot Hunter is a computational system for large-scale screening and selection of candidate immunological hotspots in pathogen proteomes (Zhang et al., 2008). It allows screening and selection of hotspots specific to HLA-A2, -A3, -B7, and -DR supertype. Prediction of peptide binding to HLA molecules has been extensively studied. Peptide binding prediction to HLA class I was shown to be highly accurate for a number of HLA class I alleles (Lin et al., 2008a; Zhang et al., in press). Predictors of peptide binding to HLA class II molecules are less accurate than those for class I. However these prediction systems are improving over time and have acceptable accuracy for peptide pre-screening (Lin et al., 2008b). Recent benchmarking has shown the best performing systems for individual HLA predictions are NetMCHpan systems (Lundegaard et al., 2010).

Vaccine development requires multiple analyses ranging from predictions that involve a single HLA molecule and a single protein, to population-based studies where multiple sequence variants are analyzed for peptides that bind multiple HLA alleles. Computational vaccine development systems require handling of extended input information, such as multiple target antigens, multiple HLA predictors, and population properties. New visualization tools are needed to enable summarization of complex data required for vaccine development. Based on our previous developments - MULTPRED, PEPVAC, and NetMHCpan, we developed MULTIPRED2, a web-based system for prediction of peptide binding to products of individual HLA alleles, combination of alleles, and supertypes. MULTIPRED2 predicts promiscuous binders of 13 HLA class I supertypes (A1, A2, A3, A24, A26, B7, B8, B27, B44, B58, B62, C1, and C4) and 13 HLA Class II supertypes (DR1, DR3, DR4, DR6, DR7, DR8, DR9, DR11, DR12, DR13, DR14, DR15, and DR16). Supertype predictions utilize predefined combinations of HLA alleles. For more specific analysis, the user can input a single allele, or any combination of HLA alleles (such as an HLA haplotype, or an individual's genotype) and perform binding prediction for these selections. MULTIPRED2 integrates prediction results from multiple HLA supertypes and displays summary views of immunological hotspots. Heat map visualization is employed to show the global view of peptide binding affinities for many HLA molecules. A heat map is a two-dimensional representation of values in a data matrix coded as colors and shades. Combined with cluster analysis heat maps can elucidate fundamental patterns in complex data. The MULTIPRED2 heat map tool enables the visualization of predicted peptide binding affinities across the input protein and multiple HLA alleles. Another useful function of MULTIPRED2 is the calculation of population coverage for selected HLA alleles. This analysis has implication in determining the proportion of population for which the selected peptides are relevant as vaccine targets.

2. Materials and methods

2.1. Selection of the prediction engines

Multiple servers are publicly available online for prediction of peptide binding to HLA class I and class II molecules. However, the lack of standardized methodology and large number of human MHC molecules make the selection of appropriate prediction servers difficult. Previously we performed comparative evaluation of 30 prediction servers for seven HLA class I molecules, HLA-A*0201, A*0301, A*1101, B*0702, B*0801, B*1501 and A*2402 (Lin et al., 2008a); and of 21 prediction servers for seven HLA-DR molecules, DRB1*0101, 0301, 0401, 0701, 1101, 1301, and 1501 (Lin et al., 2008b). Other groups also performed evaluations of prediction of peptide binding to HLA class I alleles (Gowthaman et al., 2010; Zhang et al., 2009) and their results were concordant with ours. Considering the evaluation of prediction accuracy and the number of prediction models provided by the online systems, NetMCHpan 2.0 (Nielsen et al., 2007) and NetMHCII-pan 1.0 (Nielsen et al., 2008) were selected as the prediction engines of MULTIPRED2.

2.2. Definition of HLA supertypes

HLA genes are the most polymorphic human genes; HLA polymorphism must be taken into account in design of epitope-based vaccines. HLA alleles that show similar peptide binding specificity are grouped into supertypes (Sidney et al., 1996). The majority of HLA class I alleles cluster into nine HLA class I supertypes, A1, A2, A3, A24, B7, B27, B44, B58, and B62, largely based on their overlapping peptide binding repertoires and consensus structures in the main peptide binding pockets (Sette and Sidney, 1999). The definitions of HLA-A and -B supertypes in MULTIPRED2 are predicated upon a recent extension and update of the previous classification (Sidney et al., 2008). That classification approach relies on published binding motifs, binding data, and analyses of shared repertoires of binding peptides, and the primary sequence of the B and F peptide binding pockets. Four rules for allele assignment to supertypes were reported [24]. A sequence will be assigned to a supertype if it shares patterns with other members of that supertype, including (a) presence of experimentally established motifs, (b) exact matches of all residues in B and F pockets of the HLA groove, (c) one exact match for all residues and one exact match of key residues in pockets B and F, or (d) exact matches of key residues in B and F pockets. In MULTIPRED2 we only used allele assignments based on experimentally established motif or exact matches in the B and F pockets. One more HLA-A supertype, A26, has been included in MULTIPRED2. The definition of A26 alleles was based on the analysis of specificity matrices (Lund et al., 2004). The definition of HLA-C supertypes are based on structural similarities and molecular interaction fields calculated for the peptide binding sites (Doytchinova et al., 2004). The definition of HLA-DR supertypes in MULTIPRED2 are based on classification from the HLA dictionary (Holdsworth et al., 2009). The 671 alleles belonging to the 13 HLA class I supertypes are listed at http://cvc.dfci.harvard.edu/multipred2/HTML/reference.php#Q1. The 406 alleles belonging to the 13 HLA class II supertypes are listed at http://cvc.dfci.harvard.edu/multipred2/HTML/reference.php#Q2.

2.3. Calculation of population coverage

Each of the A2, A3, B44, and B7 supertypes covers 35–55% of the general population (Sidney et al., 1996). The A2, A3 and B7 supertypes together cover ≥83% of the human population, while the A1, A2, A3, A24, B7 and B44 supertypes collectively cover ≥98% of the human population (Sette and Sidney, 1999). HLA-A, B, and C loci were typed at the allele level using PCR-based methods in 1296 unrelated subjects from five major groups living in the USA, African American; Caucasians; Asian; Hispanic, and North American Natives (Cao et al., 2001). These reported allele frequencies were used in the calculation of frequencies of HLA class I supertypes (Reche et al., 2006).

The cumulative phenotypic frequency (CPF) of a supertype is calculated using $CPF = 1-\left(1-\sum_{i\in A} p_i\right)^2$, assuming Hardy-Weinberg proportions for the genotypes (Dawson et al., 2001), where p_i is the population frequency of the i^{th} alleles within a supertype A. The CPF of a HLA-A supertype and a HLA-B supertype is calculated using $CPF = 1-\left(1-\sum_{i\in A} p_i-\sum_{j\in B} q_j+\sum_{i\in A}\sum_{j\in B} h_{ij}\right)^2$, where p_i and q_j are, respectively, the population frequencies of the ith HLA-A allele within the supertype A and the jth HLA-B allele within the supertype B, and h_{ij} denotes the haplotype frequency for the ith HLA-A and jth HLA-B variants.

2.4. Pre-calculated representative viral proteomes

There are 11 representative viral proteomes, covering five viral species, with their binding prediction pre-calculated and stored in MULTIPRED2. They include two west Nile virus strains - 956 (Genbank accession: NP_041724) and NY99-flamingo382-99 (Genbank accession: AAF20092); a yellow fever virus 17D vaccine strain (Genbank accession: NP_041726), a SARS corona virus (Genbank accession: NP_828849), four serotypes of dengue viruses, (Genbank accessions: ABW82089, ACA48914, ABW82024, ACW82884); and three influenza A virus strains - Influenza A/Goose/Guangdong/1/1996(H5N1), Influenza A/Mexico/4108/2009 (H1N1), and Influenza A/Brevig Mission/1/1918(H1N1). We broke down the influenza A virus proteomes into individual proteins — 11 proteins for Influenza A/Goose/Guangdong/1/1996(H5N1) and Influenza A/Brevig Mission/1/1918(H1N1), and 10 proteins for Influenza A/Mexico/4108/2009(H1N1) because 4108 does not have PB1-F2 protein. PB1-F2, a product of an alternative reading frame in the PB1-encoding RNA segment 2, is a key danger factor which distinguishes the major flu pandemics of the 20th century (1918 Spanish, 1957 Asian, and 1968 Hong Kong) from the milder 2009 H1N1 “swine flu” pandemic. The details of the three influenza A virus proteins are accessible at http://cvc.dfci.harvard.edu/multipred2/HTML/sequence.php.

Previously, we performed a large-scale analysis of the evolutionary variability of the influenza A virus proteins (Heiny et al., 2007). The sequence diversity and conservation study was performed on 36,343 sequences of the 11 viral proteins of human H1N1, H3N2, H1N2, H5N1, avian H5N1, and other avian subtypes circulating between 1997 and 2006. Fifty-five highly conserved sequences, which are conserved in at least 80% or more of the protein sequences of the analyzed dataset, were identified. In the immunological hotspot display page, these conserved sequences are shown in bold italic letters as shown in Fig. 2.

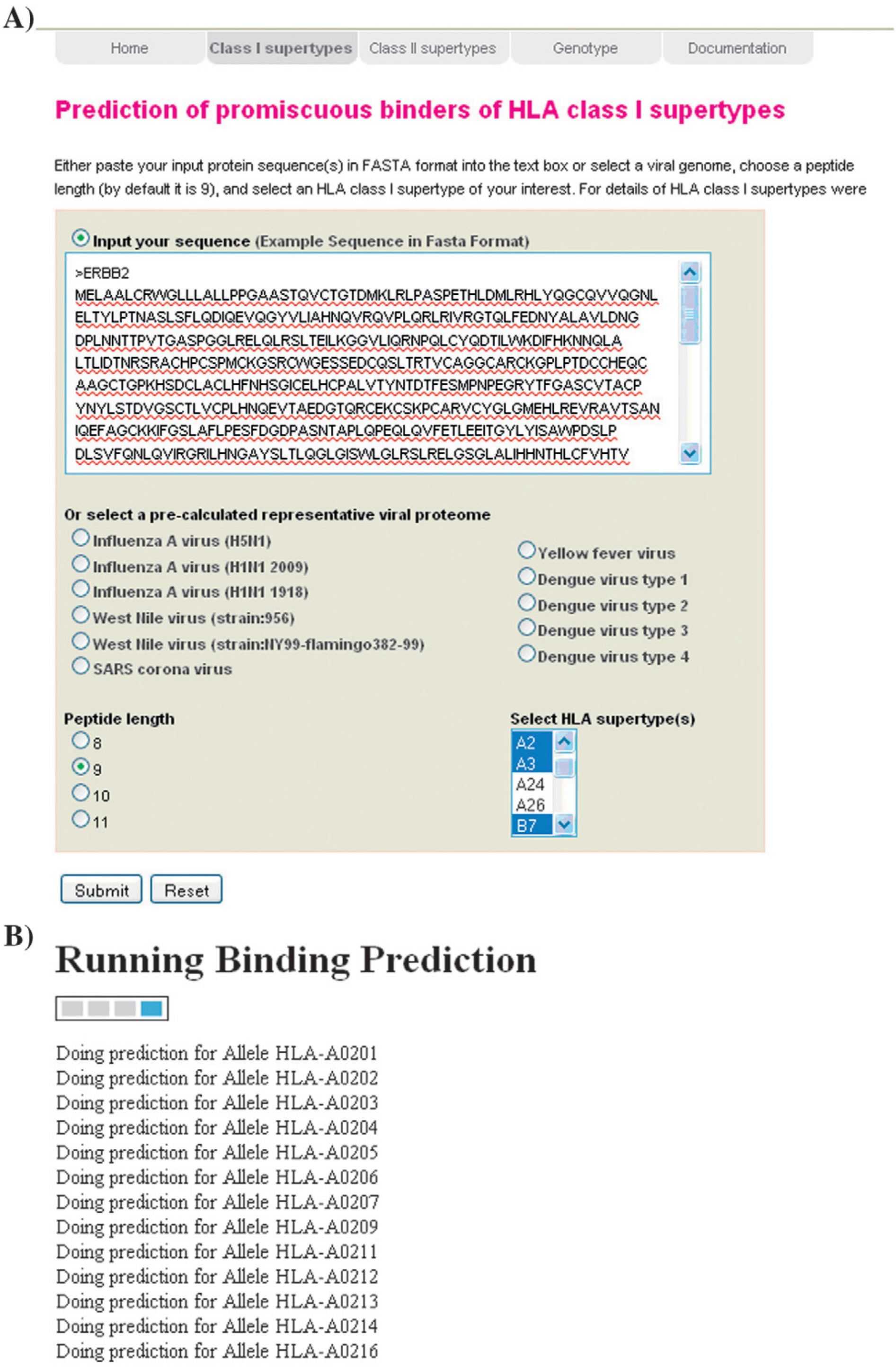
2.5. Heat maps

Heat maps are generated using the GenePattern analysis platform (Reich et al., 2006). Data is first prepared in the GCT format. IC_{50} affinity scores are log-transformed into a linear scale which is more appropriate for heat map shading. The GenePattern Hierarchical Clustering module is used to cluster alleles within a supertype according to their binding patterns across all peptides (de Hoon et al., 2004). The Pearson correlation distance measure is used for clustering HLA alleles together by their peptide binding preferences. When clustering peptides together by HLA specificity, the Euclidean distance is applied to avoid clustering weak binders together with strong binders. The HeatMapImage module creates the heat map as a JPEG image file. For this module the color scheme was set to global to ensure that shading corresponds directly with binding affinity. Network interfacing with GenePattern is achieved using the GenePattern java API. Integration into the MULTIPRED2 web interface was implemented in Java Server Pages hosted on an Apache Tomcat Server.

3. Using the system

The web interface of MULTIPRED2 uses a set of Graphical User Interface forms with a combination of Perl, Java, and C background programs. Development of MULTIPRED2 was carried out in CentOS Linux environment. The functions provided by MULTIPRED2 include (1) predicting promiscuous binders for 13 HLA class I and 13 class II supertypes; (2) predicting binders specific to an individual; (3) calculating population coverage in five ethnic groups in North America for user-selected combination of supertypes; (4) displaying immunological hotspots in an input protein; 5) visualizing global binding patterns using heat maps.

To identify promiscuous binding peptides in input protein sequences, users first go to "Class I Supertype" or "Class II Supertype" page. Then they select the radio button "Input your sequence", which is the default selection, and paste one or more FASTA format protein sequences in the text box. The next step is the selection of peptide length (8 to 11, the default length is 9) and selection of one or more supertypes of interest. The final step is clicking on the "Submit" button.



Prediction of promiscuous binders of HLA class I supertypes

Either paste your input protein sequence(s) in FASTA format into the text box or select a viral genome, choose a peptide length (by default it is 9), and select an HLA class I supertype of your interest. For details of HLA class I supertypes were

● Input your sequence (Example Sequence in Fasta Format)

```
>ERBB2
MELAALCRWGLLLALLPPGAASTQVCTGTDMKLRLPASPETHLDMLRHLYQGCQVVQGNL
ELTYLPTNASLSFLQDIQEVQGYVLIAHNQVRQVPLQRLRIVRGTQLFEDNYALAVLDNG
DPLNNTTPVTGASPGGLRELQLRSLTEILKGGVLIQRNPQLCYQDTILWKDIFHKNNQLA
LTLIDTNRSRACHPCSPMCKGSRCWGESSEDCQSLTRTVCAGGCARCKGPLPTDCCHEQC
AAGCTGPKHSDCLACLHFNHSGICELHCPALVTYNTDTFESMPNPEGRYTFGASCVTACP
YNYLSTDVGSCTLVCPLHNQEVTAEDGTQRCEKCSKPCARVCYGLGMEHLREVRAVTSAN
IQEFAGCKKIFGSLAFLPESFDGDPASNTAPLQPEQLQVFETLEEITGYLYISAWPDSLP
DLSVFQNLQVIRGRILHNGAYSLTLQGLGISWLGLRSLRELGSGLALIHHNTHLCFVHTV
```

Or select a pre-calculated representative viral proteome

○ Influenza A virus (H5N1)
○ Influenza A virus (H1N1 2009)
○ Influenza A virus (H1N1 1918)
○ West Nile virus (strain:956)
○ West Nile virus (strain:NY99-flamingo382-99)
○ SARS corona virus
○ Yellow fever virus
○ Dengue virus type 1
○ Dengue virus type 2
○ Dengue virus type 3
○ Dengue virus type 4

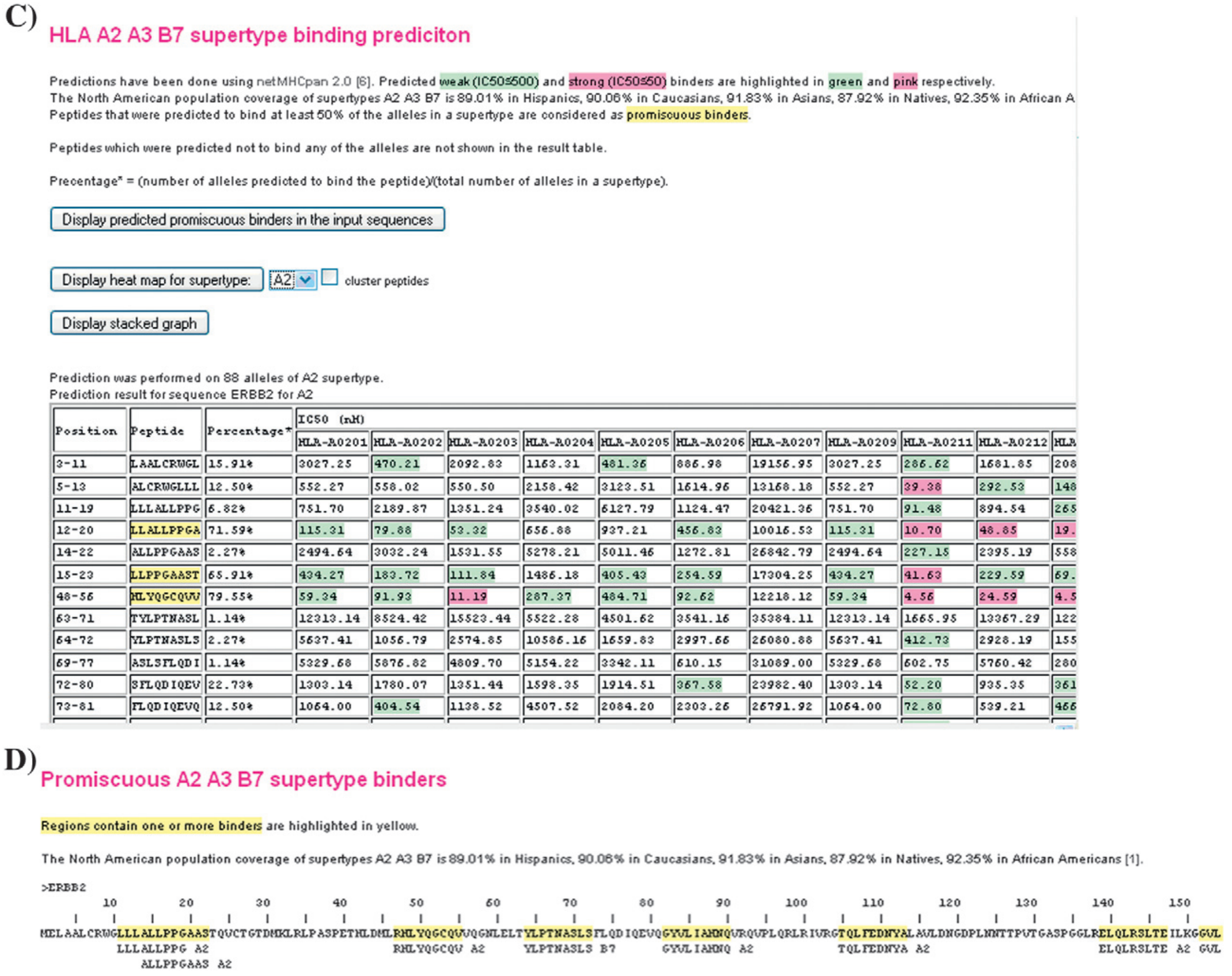
Peptide length
○ 8
● 9
○ 10
○ 11

Select HLA supertype(s)
[✓ A2]
[✓ A3]
A24
A26
[✓ B7]

Running Binding Prediction

```
Doing prediction for Allele HLA-A0201
Doing prediction for Allele HLA-A0202
Doing prediction for Allele HLA-A0203
Doing prediction for Allele HLA-A0204
Doing prediction for Allele HLA-A0205
Doing prediction for Allele HLA-A0206
Doing prediction for Allele HLA-A0207
Doing prediction for Allele HLA-A0209
Doing prediction for Allele HLA-A0211
Doing prediction for Allele HLA-A0212
Doing prediction for Allele HLA-A0213
Doing prediction for Allele HLA-A0214
Doing prediction for Allele HLA-A0216
```

Fig. 1. A series of screenshots when doing HLA class I supertype prediction on a user input protein, tumor antigen ERBB2. (A) The input page. (B) The progress page to keep user informed about the prediction progress. (C) The prediction result page. (D) The global view of immunological hotspots locations. (E) Heat map showing the global view of peptide binding profile. Each column represents the binding profile of multiple peptides to an HLA allele. HLA alleles with similar binding profiles are cluster together. Each row represents the binding profile of a peptide to multiple HLA alleles. Peptides are shown in sequential order of their position in the protein. (F) Heat map with peptide clustered based on similarity of their binding profiles.


C)
HLA A2 A3 B7 supertype binding prediciton
Display predicted promiscuous binders in the input sequences
Display heat map for supertype:
A2
□ cluster peptides
Display stacked graph
D)
Promiscuous A2 A3 B7 supertype binders

Predictions have been done using netMHCpan 2.0 [6]. Predicted weak (IC50≤500) and strong (IC50≤50) binders are highlighted in green and pink respectively.
The North American population coverage of supertypes A2 A3 B7 is 89.01% in Hispanics, 90.06% in Caucasians, 91.83% in Asians, 87.92% in Natives, 92.35% in African A
Peptides that were predicted to bind at least 50% of the alleles in a supertype are considered as promiscuous binders.

Peptides which were predicted not to bind any of the alleles are not shown in the result table.

Precentage* = (number of alleles predicted to bind the peptide)/(total number of alleles in a supertype).

Prediction was performed on 88 alleles of A2 supertype.
Prediction result for sequence ERBB2 for A2

Position	Peptide	Percentage*	IC50 (nM)										
			HLA-A0201	HLA-A0202	HLA-A0203	HLA-A0204	HLA-A0205	HLA-A0206	HLA-A0207	HLA-A0209	HLA-A0211	HLA-A0212	HLA
3-11	LAALCRWGL	15.91%	3027.25	470.21	2092.83	1163.31	481.36	886.98	19156.95	3027.25	286.62	1681.85	208
5-13	ALCRWGLLL	12.50%	552.27	558.02	550.50	2158.42	3123.51	1614.96	13168.18	552.27	39.38	292.53	148
11-19	LLLALLPPG	6.82%	751.70	2189.87	1351.24	3540.02	6127.79	1124.47	20421.36	751.70	91.48	894.54	265
12-20	LLALLPPGA	71.59%	115.31	79.88	53.32	656.88	937.21	456.83	10016.53	115.31	10.70	48.85	19.
14-22	ALLPPGAAS	2.27%	2494.64	3032.24	1531.55	5278.21	5011.46	1272.81	26842.79	2494.64	227.15	2395.19	558
15-23	LLPPGAAST	65.91%	434.27	183.72	111.84	1486.18	405.43	254.59	17304.25	434.27	41.63	229.59	69.
48-56	HLYQGCQVV	79.55%	59.34	91.93	11.19	287.37	484.71	92.62	12218.12	59.34	4.56	24.59	4.5
63-71	TYLPTNASL	1.14%	12313.14	8524.42	15523.44	5522.28	4501.62	3541.16	35384.11	12313.14	1665.95	13367.29	122
64-72	YLPTNASLS	2.27%	5637.41	1056.79	2574.85	10586.16	1659.83	2997.66	26080.88	5637.41	412.73	2928.19	155
69-77	ASLSFLQDI	1.14%	5329.68	5876.82	4809.70	5154.22	3342.11	610.15	31089.00	5329.68	602.75	5760.42	280
72-80	SFLQDIQEV	22.73%	1303.14	1780.07	1351.44	1598.35	1914.51	367.58	23982.40	1303.14	52.20	935.35	361
73-81	FLQDIQEVQ	12.50%	1064.00	404.54	1138.52	4507.52	2084.20	2303.26	26791.92	1064.00	72.80	539.21	466

Regions contain one or more binders are highlighted in yellow.

The North American population coverage of supertypes A2 A3 B7 is 89.01% in Hispanics, 90.06% in Caucasians, 91.83% in Asians, 87.92% in Natives, 92.35% in African Americans [1].

```
>ERBB2
         10        20        30        40        50        60        70        80        90       100       110       120       130       140       150
    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
MELAALCRWGLLLALLPPGAASTQVCTGTDMKLRLPASPETHLDMLRHLYQGCQVVQGNLELTYLPTNASLSFLQDIQEVQGYVLIAHNQVRQVPLQRLRIVRGTQLFEDNYALAVLDNGDPLNNTTPVTGASPGGLRELQLRSLTEILKGGVL
          LLLALLPPG A2                               RHLYQGCQV A2     YLPTNASLS B7        GYVLIAHNQ A2              TQLFEDNYA A2                            ELQLRSLTE A2 GVL
             ALLPPGAAS A2
```

Fig. 1 (*continued*).

Fig. 1A is a screenshot of a class I prediction input page, in which the input includes a FASTA format protein sequence pasted in the text box, the selected peptide length 9, and three selected supertypes, A2, A3, and B7. The example case has the sequence of the ERBB2 protein, a tumor antigen, as input. The process involved predictions on 88 A2 alleles, 74 A3 alleles, and 111 B7 alleles. To keep users informed, an intermediate page (see Fig. 1B) is displayed to report the progress. When all the predictions are completed, a result page (see Fig. 1C) will automatically appear in the users' web browser.

In the result table, the first column shows the positions in the protein sequence of the 9mer peptides, the second column shows the amino acid sequences of the peptides, the third column show a percentage value. The result table is typically wider than the screen. The predicted weak ($IC_{50} \leq 500$) and strong ($IC_{50} \leq 50$) binders are highlighted in green and pink, respectively, in the result tables. Predicted promiscuous binding peptides are highlighted in yellow, which are predicted to bind at least 50% of the alleles in a supertype. Peptides that are predicted not to bind any of the alleles are not displayed in the result tables. The estimated coverage of supertypes A2, A3 and B7 in the North American population is displayed on top of the page. There are several buttons on the result page to facilitate further analyses. The selection of first button, "Display predicted promiscuous binders in the input sequences", produces a global view of the locations of the potential promiscuous T-cell epitopes (Fig. 1D). The second button is "Display heat map for supertype" and there is a drop-down menu for users to select a supertype of interest if the prediction was done on multiple supertypes. The result page displays a heat map for each input protein (Fig. 1E) giving a global view of binding affinities across the full range of alleles in the selected supertype and all the peptides in the protein. In the heat map, the binding affinity between a peptide and an HLA allele is represented by a small square filled by a certain shade of red or blue color. Strong binders are resented by bright red squares and non-binders are represented by dark blue squares. Alleles are grouped together according to their binding behavior across the protein. Selecting the checkbox "cluster peptide" next to the supertype selection box results in a heat map with peptides clustering together based upon their HLA specificity, as shown in Fig. 1F. In Fig. 1E and F, each column represents

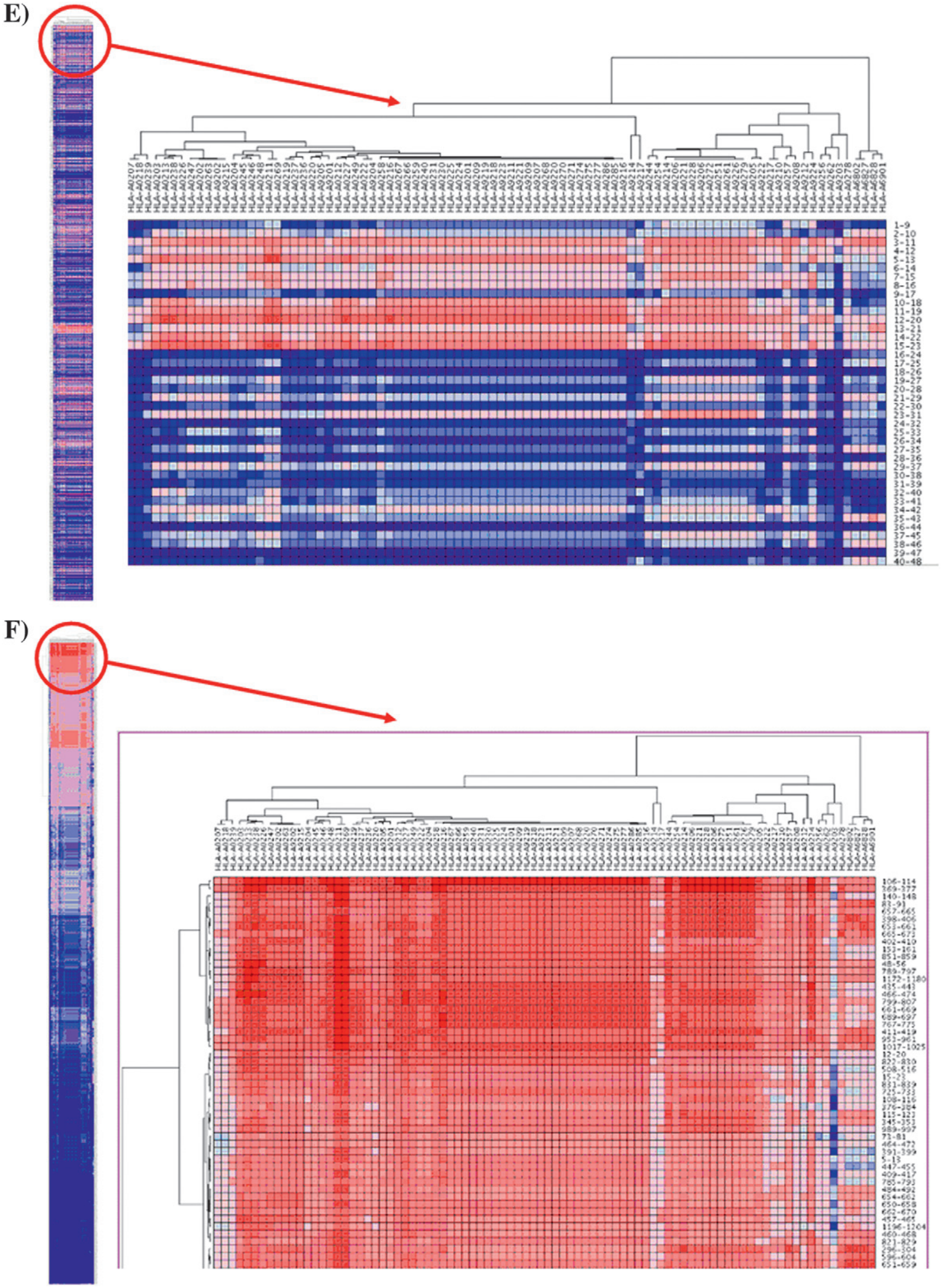


Fig. 1 (*continued*).

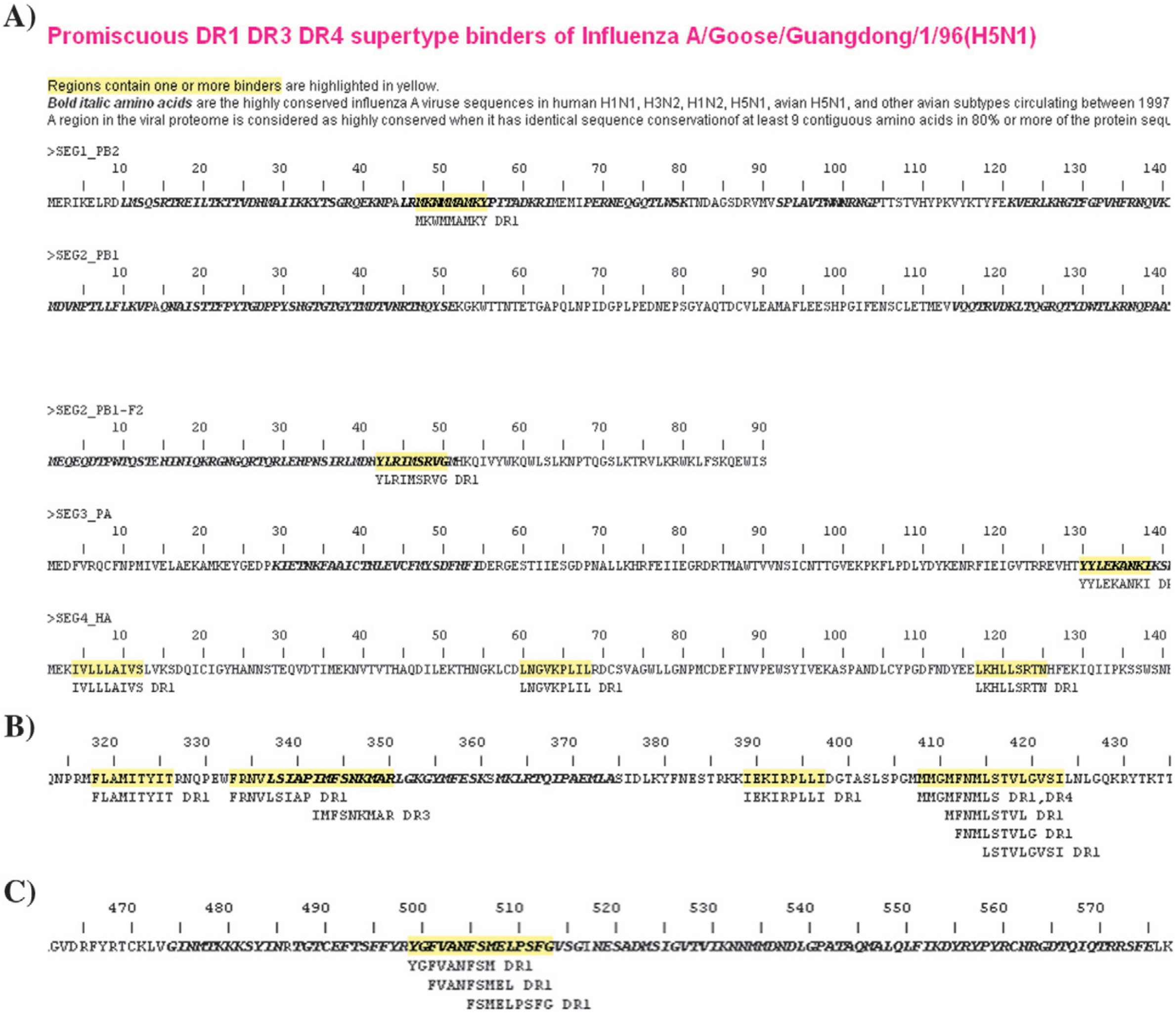
the binding profile of multiple peptides to an HLA allele, while each row represents the binding profile of a peptide to multiple HLA alleles. HLA alleles with similar binding profiles are cluster together. Peptides are shown in sequential order of their position in the protein. In Fig. 1F, instead of displaying peptides in the order of their positions, the peptides are clustered based on similarity of their HLA restriction.

To identify promiscuous binding peptides within the 11 representative viral proteomes, users first need to select the "Class I Supertype" or "Class II Supertype" page followed by selection of a radio button next to 1 of the 11 representative viral proteomes. The next step is the selection of peptide length (8–11, the default length is 9) and selection or one or more supertypes of interest. Lastly, selection of "Submit" button will produce a global view of immunological hotspots. Fig. 2 shows the result page as the global view of immunological hotspots in H5N1 influenza A virus for HLA class II DR1, DR3, and DR4 supertype prediction.

To identify binding peptides specific to a given individual in the user input protein sequences, users first select the "Genotype" page. The second step involves pasting FASTA format protein sequences (one or more) in the text box. The next step is the selection of a peptide length (8–11, the default length is 9) followed by input of a list of HLA alleles belonging to an individual genotype into the "HLA genotype" box. When all selections are completed, clicking on the "Submit" button will submit the request. This tool also allows predictions of peptide binding to a single HLA allele — by pasting this allele into the HLA genotype box.

4. Discussion

Experimental approaches for identification of T-cell epitopes are not applicable for large-scale studies involving multiple HLA alleles and pathogen proteomes as they are laborious and costly. Computational systems are widely used



A)

Promiscuous DR1 DR3 DR4 supertype binders of Influenza A/Goose/Guangdong/1/96(H5N1)

Regions contain one or more binders are highlighted in yellow.
Bold italic amino acids are the highly conserved influenza A viruse sequences in human H1N1, H3N2, H1N2, H5N1, avian H5N1, and other avian subtypes circulating between 1997
A region in the viral proteome is considered as highly conserved when it has identical sequence conservationof at least 9 contiguous amino acids in 80% or more of the protein sequ

```
>SEG1_PB2
        10        20        30        40        50        60        70        80        90       100       110       120       130       140
    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
MERIKELRDLMSQSRTREILTKTTVDHMAIIKKYTSGRQEKNPALRMKWMMAMKYPITADKRIMEMIPERNEQGQTLWSKTNDAGSDRVMVSPLAVTWWNRNGPTTSTVHYPKVYKTYFEKVERLKHGTFGPVHFRNQVK
                                              MKWMMAMKY DR1

>SEG2_PB1
        10        20        30        40        50        60        70        80        90       100       110       120       130       140
    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
MDVNPTLLFLKVPAQNAISTTFPYTGDPPYSHGTGTGYTMDTVNRTHQYSEKGKWTTNTETGAPQLNPIDGPLPEDNEPSGYAQTDCVLEAMAFLEESHPGIFENSCLETMEVVQQTRVDKLTQGRQTYDWTLKRNQPAA


>SEG2_PB1-F2
        10        20        30        40        50        60        70        80        90
    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
MEQEQDTPWTQSTEHINIQKRGNGQRTQRLEHPNSIRLMDHYLRIMSRVGMHKQIVYWKQWLSLKNPTQGSLKTRVLKRWKLFSKQEWIS
                                         YLRIMSRVG DR1

>SEG3_PA
        10        20        30        40        50        60        70        80        90       100       110       120       130       140
    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
MEDFVRQCFNPMIVELAEKAMKEYGEDPKIETNKFAAICTHLEVCFMYSDFHFIDERGESTIIESGDPNALLKHRFEIIEGRDRTMAWTVVNSICNTTGVEKPKFLPDLYDYKENRFIEIGVTRREVHTYYLEKANKIKS
                                                                                                                                    YYLEKANKI DR

>SEG4_HA
        10        20        30        40        50        60        70        80        90       100       110       120       130       140
    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
MEKIVLLLAIVSLVKSDQICIGYHANNSTEQVDTIMEKNVTVTHAQDILEKTHNGKLCDLNGVKPLILRDCSVAGWLLGNPMCDEFINVPEWSYIVEKASPANDLCYPGDFNDYEELKHLLSRTNHFEKIQIIPKSSWSN
   IVLLLAIVS DR1                                            LNGVKPLIL DR1                                             LKHLLSRTN DR1
```

B)

```
     320       330       340       350       360       370       380       390       400       410       420       430
|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
QNPRMFLAMITYITRNQPEWFRNVLSIAPIMFSNKMARLGKGYMFESKSMKLRTQIPAEMLASIDLKYFNESTRKKIEKIRPLLIDGTASLSPGMMMGMFNMLSTVLGVSILNLGQKRYTKTT
     FLAMITYIT DR1  FRNVLSIAP DR1                                     IEKIRPLLI DR1     MMGMFNMLS DR1,DR4
                         IMFSNKMAR DR3                                                      MFNMLSTVL DR1
                                                                                             FNMLSTVLG DR1
                                                                                                LSTVLGVSI DR1
```

C)

```
   470       480       490       500       510       520       530       540       550       560       570
|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
GVDRFYRTCKLVGINMSKKKSYINKTGTFEFTSFFYRYGFVANFSMELPSFGVSGINESADMSIGVTVIKNNMINNDLGPATAQMALQLFIKDYRYTYRCHRGDTQIQTRRSFELK
                                     YGFVANFSM DR1
                                       FVANFSMEL DR1
                                          FSMELPSFG DR1
```

Fig. 2. The screenshots of the global view of immunological hotspots in H5N1 influenza A virus for combined HLA class II DR1, DR3, and DR4 supertype prediction. The sequence in bold italic represents highly conserved residues of influenza A virus sequences in H1N1, H3N2, H1N2, and H5N1 circulating strains. The single screen view shown here accommodates 140 residues of the antigen, so most sequence displays are wider than a single screen. Therefore, we provided several views: (A) top left of the result page; (B) candidate hotspots in PB1: a DR1/DR3 candidate hotspot (333–350) and a DR1/DR4 candidate hotspot (408–423); (C) a DR1 candidate hotspot (499–513) in PB1; this hotspot is within a highly conserved region.

to complement experimental studies for identification of HLA binding peptides. In our recent study (Riemer et al., 2010), we interrogated the MHC class I peptide array of HLA-A*0201 Human Papillomavirus (HPV) 16 transformed epithelial tumor cells for the presence of HLA-A*0201-binding E6- and E7-derived peptides. Among the two proteins, 21 were predicted to bind HLA-A*0201, 10 were confirmed as binders, and a single conserved E7 9mer epitope, $E7_{11-19}$, was found by mass spectrometry to be expressed on HPV-16 transformed cells. *In Silico* prediction of potential T-cell epitopes was performed as the first screening step before HLA-A*0201 binding assay. In a final step, predictions on 116 HLA-A2 alleles indicated that $E7_{11-19}$ has the capacity to bind 100 of the 116 HLA-A2 alleles, indicating it is a suitable vaccine target across the majority of alleles within the HLA-A2 supertype. A practical implication of application of our tool is that $E7_{11-19}$ appears to be a universal HPV vaccine target across different populations irrespective of relative prevalence of HLA-A2 alleles.

MULTIPRED2 offers extended functionality needed for large-scale vaccine studies. It performs HLA binding peptide prediction at various resolutions (allele, haplotype, genotype, and supertype), peptide binding to one or more alleles of an individual's genotype, peptides binding to the majority of alleles of one HLA supertype, and peptides binding to the majority of alleles of multiple HLA supertypes. MULTIPRED2 enables binding predictions for 1077 alleles belonging to 26 HLA supertypes.

Displaying binding affinities in the form of a heat map provides the user with a global view of binding profiles across multiple HLA alleles to a given antigen, or even a complete viral proteome. The clustering feature groups HLA alleles by similarity of their binding profiles for a given antigen. This can help identify situations in which sub populations within a supertype must be targeted individually by vaccine components. Furthermore, clustering of peptides based on their HLA binding preference reveals redundancy of allelic coverage in peptide pools. Together, these visualization components inform the target selection process to maximize population coverage in peptide-based vaccines.

MULTIPRED2 is a useful tool for pre-screening of key antigenic regions to minimize the number of experiments required for mapping of promiscuous T-cell epitopes and T-cell epitope hotspots.

Acknowledgements

This work was supported by NIH grants U19 AI 57330 and U01 AI 90043 and a grant from DOD W81XWH-07-1-0080.

References

Bjorkman, P.J., Saper, M.A., Samraoui, B., Bennett, W.S., Strominger, J.L., Wiley, D.C., 1987. Structure of the human class I histocompatibility antigen, HLA-A2. Nature 329, 506.

Brown, S.A., Stambas, J., Zhan, X., Slobod, K.S., Coleclough, C., Zirkel, A., Surman, S., White, S.W., Doherty, P.C., Hurwitz, J.L., 2003. Clustering of Th cell epitopes on exposed regions of HIV envelope despite defects in antibody activity. J. Immunol. 171, 4140.

Cao, K., Hollenbach, J., Shi, X., Shi, W., Chopek, M., Fernandez-Vina, M.A., 2001. Analysis of the frequencies of HLA-A, B, and C alleles and haplotypes in the five major ethnic groups of the United States reveals high levels of diversity in these loci and contrasting distribution patterns in these populations. Hum. Immunol. 62, 1009.

Dawson, D.V., Ozgur, M., Sari, K., Ghanayem, M., Kostyu, D.D., 2001. Ramifications of HLA class I polymorphism and population genetics for vaccine development. Genet. Epidemiol. 20, 87.

de Hoon, M.J., Imoto, S., Nolan, J., Miyano, S., 2004. Open source clustering software. Bioinformatics 20, 1453.

Doytchinova, I.A., Guan, P., Flower, D.R., 2004. Identifiying human MHC supertypes using bioinformatic methods. J. Immunol. 172, 4314.

Gowthaman, U., Chodisetti, S.B., Parihar, P., Agrewala, J.N., 2010. Evaluation of different generic in silico methods for predicting HLA class I binding peptide vaccine candidates using a reverse approach. Amino Acids. 39 (5), 1333.

Gupta, V., Tabiin, T., Sun, K., Chandrasekaran, A., Anwar, A., Yang, K., Chikhlikar, P., Salmon, J., Brusic, V., Marques, E., Srinivasan, K.N., August, J.T., 2006. SARS coronavirus nucleocapsid immunodominant T-cell epitope cluster is common to both exogenous recombinant and endogenous DNA-encoded immunogens. Virology 347, 127.

Heiny, A.T., Miotto, O., Srinivasan, K.N., Khan, A.M., Zhang, G.L., Brusic, V., Tan, T.W., August, J.T., 2007. Evolutionarily conserved protein sequences of influenza a viruses, avian and human, as vaccine targets. PLoS ONE 2, e1190.

Holdsworth, R., Hurley, C.K., Marsh, S.G., Lau, M., Noreen, H.J., Kempenich, J.H., Setterholm, M., Maiers, M., 2009. The HLA dictionary 2008: a summary of HLA-A, -B, -C, -DRB1/3/4/5, and -DQB1 alleles and their association with serologically defined HLA-A, -B, -C, -DR, and -DQ antigens. Tissue Antigens 73, 95.

Kim, S.K., DeMars, R., 2001. Epitope clusters in the major outer membrane protein of *Chlamydia trachomatis*. Curr. Opin. Immunol. 13, 429.

Lin, H.H., Ray, S., Tongchusak, S., Reinherz, E.L., Brusic, V., 2008a. Evaluation of MHC class I peptide binding prediction servers: applications for vaccine research. BMC Immunol. 9, 8.

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

Lund, O., Nielsen, M., Kesmir, C., Petersen, A.G., Lundegaard, C., Worning, P., Sylvester-Hvid, C., Lamberth, K., Roder, G., Justesen, S., Buus, S., Brunak, S., 2004. Definition of supertypes for HLA molecules using clustering of specificity matrices. Immunogenetics 55, 797.

Lundegaard, C., Lund, O., Nielsen, M., 2010. Prediction of epitopes using neural network based methods. J. Immunol. Meth. (Epub ahead of print).

Meuer, S.C., Schlossman, S.F., Reinherz, E.L., 1982. Clonal analysis of human cytotoxic T lymphocytes: T4+ and T8+ effector T cells recognize products of different major histocompatibility complex regions. Proc. Natl Acad. Sci. USA 79, 4395.

Nielsen, M., Lundegaard, C., Blicher, T., Lamberth, K., Harndahl, M., Justesen, S., Roder, G., Peters, B., Sette, A., Lund, O., Buus, S., 2007. NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. PLoS ONE 2, e796.

Nielsen, M., Lundegaard, C., Blicher, T., Peters, B., Sette, A., Justesen, S., Buus, S., Lund, O., 2008. Quantitative predictions of peptide binding to any HLA-DR molecule of known sequence: NetMHCIIpan. PLoS Comput. Biol. 4, e1000107.

Reche, P.A., Keskin, D.B., Hussey, R.E., Ancuta, P., Gabuzda, D., Reinherz, E.L., 2006. Elicitation from virus-naive individuals of cytotoxic T lymphocytes directed against conserved HIV-1 epitopes. Med Immunol 5, 1.

Reche, P.A., Reinherz, E.L., 2005. PEPVAC: a web server for multi-epitope vaccine development based on the prediction of supertypic MHC ligands. Nucleic Acids Res. 33, W138.

Reich, M., Liefeld, T., Gould, J., Lerner, J., Tamayo, P., Mesirov, J.P., 2006. GenePattern 2.0. Nat. Genet. 38, 500.

Riemer, A.B., Keskin, D.B., Zhang, G., Handley, M., Anderson, K.S., Brusic, V., Reinhold, B., Reinherz, E.L., 2010. A Conserved E7-derived Cytotoxic T Lymphocyte Epitope Expressed on Human Papillomavirus 16-transformed HLA-A2+ Epithelial Cancers. J. Biol. Chem. 285, 29608.

Robinson, J., Waller, M.J., Fail, S.C., McWilliam, H., Lopez, R., Parham, P., Marsh, S.G., 2009. The IMGT/HLA database. Nucleic Acids Res. 37, D1013.

Sette, A., Sidney, J., 1999. Nine major HLA class I supertypes account for the vast preponderance of HLA-A and -B polymorphism. Immunogenetics 50, 201.

Sidney, J., Grey, H.M., Kubo, R.T., Sette, A., 1996. Practical, biochemical and evolutionary implications of the discovery of HLA class I supermotifs. Immunol. Today 17, 261.

Sidney, J., Peters, B., Frahm, N., Brander, C., Sette, A., 2008. HLA class I supertypes: a revised and updated classification. BMC Immunol. 9, 1.

Surman, S., Lockey, T.D., Slobod, K.S., Jones, B., Riberdy, J.M., White, S.W., Doherty, P.C., Hurwitz, J.L., 2001. Localization of CD4+ T cell epitope hotspots to exposed strands of HIV envelope glycoprotein suggests structural influences on antigen processing. Proc. Natl Acad. Sci. USA 98, 4587.

Wang, J.H., Reinherz, E.L., 2002. Structural basis of T cell recognition of peptides bound to MHC molecules. Mol. Immunol. 38, 1039.

Zhang, G.L., Bozic, I., Kwoh, C.K., August, J.T., Brusic, V., 2007. Prediction of supertype-specific HLA class I binding peptides using support vector machines. J Immunol Methods 320, 143.

Zhang, G.L., et al., in press. Machine learning in immunology competition: prediction of HLA class I ligands. Journal of Immunological Methods.

Zhang, G.L., Khan, A.M., Srinivasan, K.N., August, J.T., Brusic, V., 2005. MULTIPRED: a computational system for prediction of promiscuous HLA binding peptides. Nucleic Acids Res 33, W172.

Zhang, G.L., Khan, A.M., Srinivasan, K.N., Heiny, A., Lee, K., Kwoh, C.K., August, J.T., Brusic, V., 2008. Hotspot Hunter: a computational system for large-scale screening and selection of candidate immunological hotspots in pathogen proteomes. BMC Bioinformatics 9 (Suppl 1), S19.

Zhang, H., Lundegaard, C., Nielsen, M., 2009. Pan-specific MHC class I predictors: a benchmark of HLA class I pan-specific prediction methods. Bioinformatics 25, 83.