W172–W179 *Nucleic Acids Research, 2005, Vol. 33, Web Server issue*
doi:10.1093/nar/gki452

MULTIPRED: a computational system for prediction of promiscuous HLA binding peptides

Guang Lan Zhang[1,2], Asif M. Khan[1,3], Kellathur N. Srinivasan[4,5], J. Thomas August[4,5] and Vladimir Brusic[1,6,*]

[1]Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613, [2]School of Computer Engineering, Nanyang Technological University, Singapore 639798, [3]Department of Biochemistry, National University of Singapore, Singapore 117597, [4]Department of Pharmacology and Molecular Sciences, Johns Hopkins School of Medicine, Baltimore, MD 21205, USA, [5]Division of Biomedical Sciences, Johns Hopkins in Singapore, #02-01 The Nanos, 31 Biopolis Way, Singapore 138669 and [6]School of Land and Food Sciences and the Institute for Molecular Bioscience, University of Queensland, Brisbane QLD 4072, Australia

Received February 14, 2005; Revised and Accepted April 1, 2005

ABSTRACT

MULTIPRED is a web-based computational system for the prediction of peptide binding to multiple molecules (proteins) belonging to human leukocyte antigens (HLA) class I A2, A3 and class II DR supertypes. It uses hidden Markov models and artificial neural network methods as predictive engines. A novel data representation method enables MULTIPRED to predict peptides that promiscuously bind multiple HLA alleles within one HLA supertype. Extensive testing was performed for validation of the prediction models. Testing results show that MULTIPRED is both sensitive and specific and it has good predictive ability (area under the receiver operating characteristic curve $A_{ROC} > 0.80$). MULTIPRED can be used for the mapping of promiscuous T-cell epitopes as well as the regions of high concentration of these targets—termed T-cell epitope hotspots. MULTIPRED is available at http://antigen.i2r.a-star.edu.sg/multipred/.

INTRODUCTION

T-cells of the human immune system recognize antigens as short peptide fragments (T-cell epitopes) derived from the degradation of proteins. Major histocompatibility complex (MHC) proteins play a vital role in the initiation and regulation of immune responses (1–4). Their primary function is to bind and subsequently present antigenic peptides on the cell surface for recognition by T-cells of the immune system. The recognition of T-cell epitopes is critical for the immune response to infectious, autoimmune, allergic and neoplastic disease. T-cell epitopes are important for the development of peptide-based vaccines (5). There is a great diversity of human leukocyte antigens (HLAs; human MHC) genes with some 2000 known variants characterized to date (6). HLA proteins share 3D structure with main differences observed in residues that form the peptide-binding groove. HLA proteins that have small differences in their peptide-binding grooves and share similar peptide-binding specificities are grouped into HLA supertypes (7,8). Promiscuous peptides—those that bind more than one HLA variant—are prime targets for vaccine and immunotherapy development because they are relevant to higher proportions of the human population. Because of the large number of HLA proteins, experimental approaches for identifying T-cell epitopes (from overlapping peptides that span the length of protein antigens) are time-consuming and costly, and thus not applicable for large-scale screening. Computer modeling methods can help to simulate the biological process of antigen presentation, minimize the number of experiments required, enable a systematic scanning for candidate MHC-binding peptides and thus speed up vaccine development (9).

MULTIPRED is a web-based system for the prediction of peptides that bind multiple HLA alleles. Current implementation can predict peptides that bind HLA proteins belonging to supertypes A2 and A3 (HLA class I) as well as DR (HLA class II) and in future will be extended to other supertypes. The predictive engines implemented in MULTIPRED are hidden Markov models (HMMs) and artificial neural networks (ANNs). A novel data representation method enables MULTIPRED to predict peptides that bind to multiple HLA alleles belonging to one HLA supertype by a single prediction model per supertype.

SYSTEM DESCRIPTION

The predominant length of peptides that bind HLA-A2 and -A3 (class I) proteins is nine amino acids (10). HLA-DR

*To whom correspondence should be addressed. Tel: +65 96212 415; Fax: +65 6774 8056; Email: vladimir@i2r.a-star.edu.sg

(class II) proteins bind longer peptides through the core binding region, which is nine amino acids long (11). The training data comprise 3050 9mer peptide sequences (664 binders and 2386 non-binders) related to 15 variants of the HLA-A2 supertype, 2216 9mer peptide sequences (680 binders and 1536 non-binders) related to eight variants of the HLA-A3 supertype and 2396 9mer peptides (448 binders and 1948 non-binders) related to six HLA-DR variants. These data are mainly from three sources, the MHCPEP database (12), published articles and a set of HLA non-binding peptides (V. Brusic, unpublished data). For both training and prediction the data representation includes both the peptide and its binding environment (HLA contact residues). This 'virtual peptide' representation comprises both peptide residues and the environment for each residue of the 9mer peptides (13,14). To simplify the data representation and eliminate redundant information, for each HLA supertype, we considered only those contact residues that vary across various HLA variants and discarded the residues, which are conserved.

In MULTIPRED, a three-layer backpropagation network with sigmoid activation functions was built for HLA-A2 and -A3 supertype and a four-layer backpropagation network with a hyperbolic tangent sigmoid activation function between the two hidden layers and a sigmoid activation function between the second hidden layer and the output for HLA-DR supertype. Various techniques, including optimization of ANN architecture and balancing datasets, were explored to improve the prediction accuracy of the ANN models (14). MULTIPRED also has a first-order HMM as an alternative prediction engine (13). The user can select either the ANN or the HMM model for prediction—both methods have been optimized and show similar performance. The A_{ROC} is >0.8 in all cases, indicating good prediction capability [see (13,14) for details on HLA-A2 models, (15) for HLA -A3 models, and V. Brusic, A. Sette, G. L. Zhang, K. N. Srinivasan, J. T. August and V. Brusic, manuscript in preparation for HLA-DR models.

In addition to individual 9mer predictions, MULTIPRED also predicts immunological hotspots (regions of high concentration of 9mer promiscuous binders). We have developed two scoring schemes to identify immunological hotspots within antigens for HLA classes I and II supertype. The scheme for HLA class I supertype is based on high-scoring individual 9mers within a window of 30 amino acids (15) and the scheme for HLA class II supertype is based on average scores of individual 9mers within a window of 15 amino acids. The selection of window lengths was based on a trial-and-error process. Window lengths of 15, 20, 25 and 30, were explored and the results were compared with the representative experimental results. The window length 30 was found to suit class I predictions and window length 15 to class II predictions. The lengths outside these ranges are considered too short or too long as targets for experimental validation. The prediction performance of MULTIPRED for HLA-A2 and -A3 hotspots was validated using experimental results from a systematic study of human papillomavirus type 16 E6 (P03126) and E7 (P03129) proteins (16). The prediction performance of MULTIPRED for HLA-DR hotspots was validated using experimental results from systematic binding studies of overlapping peptides from Myelin Oligodendrocyte glycoprotein (MOG) (CAA88109), bee venom protein (1POC) and hepatitis C virus 1B protein (AAB00216).

USING THE SYSTEM

The web interface of MULTIPRED uses a set of graphical user interface forms with a combination of Perl, CGI and C background programs. Development of MULTIPRED was carried out in SunOS 5.9 UNIX environment. The functions provided by MULTIPRED include (i) running predictions, (ii) model building, (iii) prediction accuracy evaluation and (iv) identifying consensus predictions among up to three sets of predictions on the same input protein sequence.

To predict peptides binding to a supertype, users must first select 'Run prediction'. The required input is the selection of supertype and prediction method (pre-defined ANN or HMM). Alternatively, users can select a pre-defined model (built by model building function). By selecting the 'Submit' button users get to a sequence input page where the required input is a protein sequence and its name. The length of the input sequence must be between 9 and 2000 amino acids. If the input sequence contains symbols other than amino acids (space and carriage returns are allowed) or if the sequence is outside the length limits, an error message will be displayed. The input can either be a protein sequence or a list of peptides. The default selection on the webpage is 'Protein sequence', which means the input sequence is treated as one single protein sequence and carriage returns are ignored. If users changed the sequence type to 'a list of peptide sequences', then sequences divided by carriage returns are treated as separate peptides. The processing steps and result pages for the two types of inputs are different. The detailed description on processing steps involved when the input sequence is a protein sequence or a list of peptides are available at http://antigen.i2r.a-star.edu.sg/multipred/HTML/faq.html#Q3 and http://antigen.i2r.a-star.edu.sg/multipred/HTML/faq.html#Q4, respectively. The 9mer binding scores range from 1 to 9 (Figure 1A), with scores 4–9 referring to predicted binders (8 or 9 referring to high, 6 or 7 to moderate, and 4 or 5 to low confidence of peptide binding). Scores 1–3 refer to predicted non-binders. MULTIPRED saves the prediction result and the users may note down the ID number of the saved jobs for the comparison of prediction results generated by different prediction models (Figure 1A). Two scoring schemes to identify immunological hotspots within antigens were developed for HLA classes I and II supertype. The scheme for HLA class I supertypes is based on high-scoring individual 9mers within a window of 30 amino acids (15). In the result table (Figure 1A), 'Sum' is the sum total of the individual binding scores of a peptide to the MHC proteins, 'Score 1' is the top 1 'Sum' in a 30mer window (A 30mer window comprises 22 consecutive 9mer peptides). 'Score 2' is the average of the top 2 'Sum' in a 30mer window. Similarly, 'Score 3', 'Score 4' and 'Score 5' are the average of the top 3, 4 and 5 'Sum', respectively, in a 30mer window. To show the user a clear view of the binding capacity of an input protein, Scores 1–5 of all 30mer peptides of the input protein can be displayed as graphs, in which *x*-axis represents the starting position of a 30mer window and the *y*-axis represents Score 1 (2/3/4/5) of the 30mer window. For example, in Figure 1B, which is the graph of Score 4 of the protein E6, the first three 30mer windows (starting at positions 1, 2 or 3) are 36.82 and the next two windows (starting at positions 4 or 5) have scores 39.50. The following 13 30mer windows (starting at positions 6–19) have scores >42, the recommended

threshold for Score 4 for HLA-A2 ANN models (Figure 1A), indicating a predicted hotspot, which corresponds to an experimentally determined HLA-A2 hotspot in E6 protein (16). To locate the individual 9mers with top binding scores in each 30mer window, the 'align' function can be used. Figure 1C shows an example of the alignment view of the top four 9mers in each 30mer window. The user can also identify hotspots at a certain threshold by using the 'Get hotspots' function (Figure 1D). The default values on the web page are the recommended thresholds for Score 4. In a HLA-DR

A

THE PREDICTED PEPTIDE BINDING AFFINITY TO HLA classIA2 using Artificial Neural Network Method

Tue Apr 5 11:22:43 2005

Note: In the table, the binding scores range form 1 to 9, with scores 4-9 referring to predicted MHC binders (High binders: 8-9; Moderate binders: 6-7; Low binders: 4-5). Scores of 1-3 refer to predicted non-binders.

Sequence Name : E6

Result ID : 3457

Result ID : 3457

Plot according to Score: 4 Plot

* A scoring scheme to identify immunological hotspots within antigens was developed. The scheme for HLA class I supertype is based on high scoring individual 9-mers within a window of 30 amino acids. For a 9-mer peptide, "Sum" is the sum total of the individual binding scores of the peptide to the MHC molecules. "Score1" is the top 1 "Sum" in a 30-mer window. "Score2" is the average of the top 2 "Sum" in a 30-mer window. Similarly, "Score3", "Score4" and "Score5" are the average of the top 3, 4 and 5 "Sum" in a 30-mer window, respectively.

Sort result according to Score: 4 Sort

Plot according to Score: 4 Plot

30-mer peptides aligned with top 4 9-mer predictions. Align

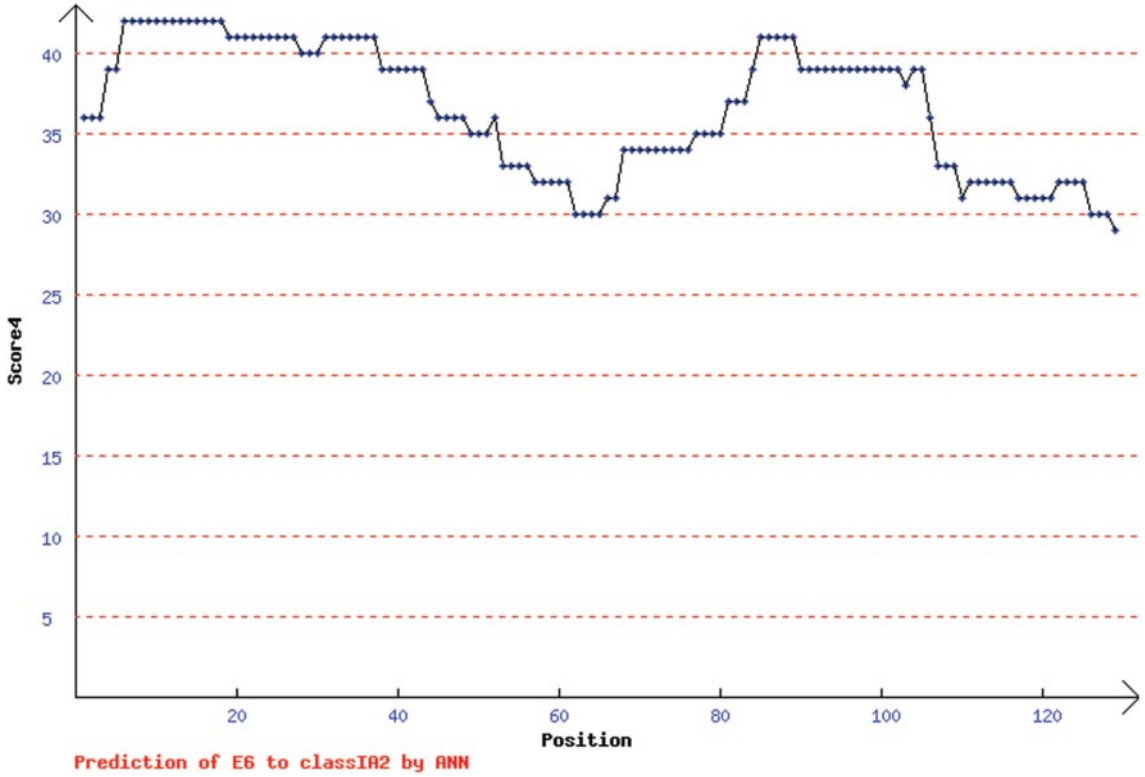
30-mer peptides aligned with top 4 9-mer predictions. Align

Using Score 4 at threshold 42 (Suggested threshold using Score4 is 42) Get hot-spots

unsorted 30-mer table

Position	Peptides	Molecules								Sum*	Score1*	Score2*	Score3*	Score4*	Score5*
		A-0201	A-0202	A-0203	A-0204	A-0205	A-0206	A-0207	A-0209						
1	MHQKRTAMF	2.61	3.21	3.26	2.63	3.66	2.80	2.65	2.61	23.43	44.94	42.92	39.05	36.82	35.06
2	HQKRTAMFQ	2.49	2.82	2.84	2.50	3.13	2.58	2.52	2.49	21.37	44.94	42.92	39.05	36.82	35.06
3	QKRTAMFQD	2.54	2.99	3.00	2.55	3.36	2.66	2.61	2.54	22.25	44.94	42.92	39.05	36.82	35.25
4	KRTAMFQDP	2.47	2.75	2.76	2.48	3.01	2.55	2.53	2.47	21.02	44.94	42.92	42.22	39.50	37.62
5	RTAMFQDPQ	2.92	3.88	3.92	2.96	4.37	3.28	2.99	2.92	27.24	44.94	42.92	42.22	39.50	37.62
6	TAMFQDPQE	2.46	2.71	2.74	2.47	2.96	2.53	2.46	2.46	20.79	44.94	43.72	42.78	42.29	40.10
7	AMFQDPQER	3.31	4.49	4.52	3.37	4.99	3.75	3.58	3.31	31.32	44.94	43.72	42.78	42.29	40.10
8	MFQDPQERP	2.44	2.62	2.64	2.44	2.82	2.49	2.45	2.44	20.34	44.94	43.72	42.78	42.29	40.12

B



C

ALIGNMENT VIEW OF THE PREDICTED BINDING TO HLA classIA2 using ANN

Sequence	Sum	Position
MHQKRTAMFQDPQERPRKLPQLCTELQTTI		1-30
KLPQLCTEL	44.94	18-26
QLCTELQTT	40.89	21-29
AMFQDPQER	31.32	7-15
ERPRKLPQL	30.13	14-22
HQKRTAMFQDPQERPRKLPQLCTELQTTIH		2-31
KLPQLCTEL	44.94	18-26
QLCTELQTT	40.89	21-29
AMFQDPQER	31.32	7-15
ERPRKLPQL	30.13	14-22
QKRTAMFQDPQERPRKLPQLCTELQTTIHD		3-32
KLPQLCTEL	44.94	18-26
QLCTELQTT	40.89	21-29
AMFQDPQER	31.32	7-15
ERPRKLPQL	30.13	14-22
KRTAMFQDPQERPRKLPQLCTELQTTIHDI		4-33
KLPQLCTEL	44.94	18-26
QLCTELQTT	40.89	21-29
ELQTTIHDI	40.84	25-33
AMFQDPQER	31.32	7-15
RTAMFQDPQERPRKLPQLCTELQTTIHDII		5-34
KLPQLCTEL	44.94	18-26
QLCTELQTT	40.89	21-29
ELQTTIHDI	40.84	25-33
AMFQDPQER	31.32	7-15
TAMFQDPQERPRKLPQLCTELQTTIHDIIL		6-35
KLPQLCTEL	44.94	18-26
QTTIHDIIL	42.50	27-35
QLCTELQTT	40.89	21-29
ELQTTIHDI	40.84	25-33
AMFQDPQERPRKLPQLCTELQTTIHDIILE		7-36
KLPQLCTEL	44.94	18-26
QTTIHDIIL	42.50	27-35

D

HLA classIA2 Hotspots predicted by ANN Method at Threshold 42 for score4

*: In the below table, "Sum" is the sum of the scores in a given row, "Score1" is the top 1 sum score in 30-mer region, "Score2" is the average of the top 2 sum score in 30-mer region, "Score3" is the average of the top 3 sum score in 30-mer region, "Score4" is the average of the top 4 sum score in 30-mer region and "Score5" is the average of the top 5 sum score in 30-mer region.

Protein Sequence Name : E6
Protein Sequence Length : 158

Position	Peptides	Score1'	Score2'	Score3'	Score4'	Score5'
6-35'	TAMFQDPQERPRKLPQLCTELQTTIHDIIL	44.94	43.72	42.78	42.29	40.10
7-36'	AMFQDPQERPRKLPQLCTELQTTIHDIILE	44.94	43.72	42.78	42.29	40.10
8-37'	MFQDPQERPRKLPQLCTELQTTIHDIILEC	44.94	43.72	42.78	42.29	40.12
9-38'	FQDPQERPRKLPQLCTELQTTIHDIILECV	44.94	43.72	42.78	42.31	42.01
10-39'	QDPQERPRKLPQLCTELQTTIHDIILECVY	44.94	43.72	42.78	42.31	42.01
11-40'	DPQERPRKLPQLCTELQTTIHDIILECVYC	44.94	43.72	42.78	42.31	42.01
12-41'	PQERPRKLPQLCTELQTTIHDIILECVYCK	44.94	43.72	42.78	42.31	42.01
13-42'	QERPRKLPQLCTELQTTIHDIILECVYCKQ	44.94	43.72	42.78	42.31	42.01
14-43'	ERPRKLPQLCTELQTTIHDIILECVYCKQQ	44.94	43.72	42.78	42.31	42.01
15-44'	RPRKLPQLCTELQTTIHDIILECVYCKQQL	44.94	43.72	42.78	42.31	42.01
16-45'	PRKLPQLCTELQTTIHDIILECVYCKQQLL	44.94	43.72	42.94	42.43	42.12
17-46'	RKLPQLCTELQTTIHDIILECVYCKQQLLR	44.94	43.72	42.94	42.43	42.12
18-47'	KLPQLCTELQTTIHDIILECVYCKQQLLRR	44.94	43.72	42.94	42.43	42.12

Predicted hotspots sorted by position:

Position	Sequences	score4'	Length
6-47'	TAMFQDPQERPRKLPQLCTELQTTIHDIILECVYCKQQLLRR	42.33	42

Predicted hotspots sorted by score4:

Rank	Position	Sequences	score4'	Length
1	6-47'	TAMFQDPQERPRKLPQLCTELQTTIHDIILECVYCKQQLLRR	42.33	42

Figure 1. An example of the output pages of MULTIPRED when the input is a single protein sequence. The input protein sequence is a human papillomavirus type 16 E6, the prediction method used is ANN and the HLA supertype of interest is HLA-A2. (**A**) The main result page. The input sequence is truncated into overlapping 9mers for the prediction of binding scores to multiple HLA-A2 variants, *0201, *0202, *0203, *0204, *0205, *0206, *0207 and *0209. The red ovals are added by the authors for the clarity of viewing. (**B**) Example graph of Score 4. (**C**) Alignment view of the top four 9mers in the 30mer windows. (**D**) The prediction hotspot region is 6–47 at threshold 42.

A **THE PREDICTED PEPTIDE BINDING AFFINITY TO HLA classIA3 using Artificial Neural Network Method**

Tue Apr 5 13:26:19 2005

Note: In the table, the binding scores range form 1 to 9, with scores 4-9 referring to predicted MHC binders (High binders: 8-9; Moderate binders: 6-7; Low binders: 4-5). Scores of 1-3 refer to predicted non-binders.
Sequence Name : HCV
Result ID : 3480

*: For a peptide, "Sum" is the sum total of the individual binding scores of the peptide to the MHC molecules.

Other Display formats of prediction result: Alignment View | Sort the Result | Plot Sum Value |

Position	Peptides	Molecules							Sum*
		A-0301	A-0302	A-1101	A-1102	A-3101	A-3301	A-6801	
1	MSTNPKEFRKTKRN	7.13	6.38	6.75	6.75	6.78	6.65	6.24	46.68
2	KEFRKTKRNTLRRP	6.84	5.88	6.16	6.16	6.61	6.42	5.55	43.62
3	TKRNTLRRPQDVRF	5.30	2.89	3.33	3.33	4.46	4.04	2.39	25.74
4	TLRRPQDVRFPGGG	5.30	2.89	3.33	3.33	4.46	4.04	2.39	25.74
5	QDVRFPGGGQIVGG	4.23	1.80	2.38	2.38	3.67	2.71	1.54	18.71
6	PGGGQIVGGVYLLP	4.69	1.93	2.58	2.58	4.00	3.10	1.79	20.67
7	QIVGGVYLLPRRGP	5.73	3.24	3.80	3.80	5.02	4.44	2.67	28.70
8	VYLLPRRGPRLGVR	6.21	4.47	4.86	4.86	5.81	5.39	4.00	35.60
9	RRGPRLGVRATRKT	6.52	4.95	5.46	5.46	5.94	5.63	4.60	38.56
10	RLGVRATRKTSERS	6.52	4.95	5.46	5.46	5.94	5.63	4.60	38.56
11	ATRKTSERSQPRGR	7.05	6.06	6.51	6.51	6.63	6.48	5.87	45.11
12	SERSQPRGRRQPIP	5.67	3.76	4.17	4.17	5.09	4.71	3.15	30.72
13	QPRGRRQPIPKARQ	5.51	3.40	3.83	3.83	4.60	4.22	2.87	28.26
14	RQPIPKARQPEGRA	5.14	2.49	3.00	3.00	4.74	3.90	2.04	24.31
15	KARQPEGRAWAQPG	4.24	2.01	2.45	2.45	3.81	3.02	1.61	19.59
16	PEGRTWAQPGYPWP	5.19	2.81	3.34	3.34	4.93	4.12	2.40	26.13
17	WAQPGYPWPLYGNE	6.03	3.66	4.25	4.25	5.70	5.07	3.10	32.06

B **THE PREDICTED BINDING TO classIA3 using ANN**

Peptide Sequence Name : HCV

Sorted Result

Rank	Position	Peptides	Molecules							Sum*
			A-0301	A-0302	A-1101	A-1102	A-3101	A-3301	A-6801	
1	1	MSTNPKEFRKTKRN	7.13	6.38	6.75	6.75	6.78	6.65	6.24	46.68
2	11	ATRKTSERSQPRGR	7.05	6.06	6.51	6.51	6.63	6.48	5.87	45.11
3	2	KEFRKTKRNTLRRP	6.84	5.88	6.16	6.16	6.61	6.42	5.55	43.62
4	10	RLGVRATRKTSERS	6.52	4.95	5.46	5.46	5.94	5.63	4.60	38.56
5	9	RRGPRLGVRATRKT	6.52	4.95	5.46	5.46	5.94	5.63	4.60	38.56
6	8	VYLLPRRGPRLGVR	6.21	4.47	4.86	4.86	5.81	5.39	4.00	35.60
7	23	PTDPRRRSRNLGKV	6.29	4.42	4.97	4.97	5.59	5.21	4.01	35.46
8	24	RRRSRNLGKVIDTL	6.29	4.42	4.97	4.97	5.59	5.21	4.01	35.46
9	22	RRPSWGPTDPRRRSR	6.10	4.14	4.57	4.57	5.69	5.19	3.59	33.85
10	17	WAQPGYPWPLYGNE	6.03	3.66	4.25	4.25	5.70	5.07	3.10	32.06
11	18	YPWPLYGNEGMGWA	5.81	3.60	4.05	4.05	5.60	4.91	3.09	31.11
12	12	SERSQPRGRRQPIP	5.67	3.76	4.17	4.17	5.09	4.71	3.15	30.72
13	7	QIVGGVYLLPRRGP	5.73	3.24	3.80	3.80	5.02	4.44	2.67	28.70
14	13	QPRGRRQPIPKARQ	5.51	3.40	3.83	3.83	4.60	4.22	2.87	28.26
15	36	LPGCSFSIFLLALL	5.71	3.03	3.61	3.61	4.81	4.29	2.47	27.53
16	37	SFSIFLLALLSCLT	5.34	3.04	3.51	3.51	4.54	4.16	2.55	26.65
17	20	GMGWAGWLLSPRGS	5.36	2.84	3.38	3.38	5.00	4.24	2.41	26.61
18	16	PEGRTWAQPGYPWP	5.19	2.81	3.34	3.34	4.93	4.12	2.40	26.13
19	32	LAHGVRVLEDGVNY	5.58	2.76	3.48	3.48	4.38	3.85	2.46	25.99
20	33	VRVLEDGVNYATGN	5.58	2.76	3.48	3.48	4.38	3.85	2.46	25.99
21	3	TKRNTLRRPQDVRF	5.30	2.89	3.33	3.33	4.46	4.04	2.39	25.74
22	4	TLRRPQDVRFPGGG	5.30	2.89	3.33	3.33	4.46	4.04	2.39	25.74
23	26	IDTLTCGFADLMGY	5.39	2.57	3.21	3.21	4.43	3.84	2.18	24.83
24	27	TCGFADLMGYIPLV	5.39	2.57	3.21	3.21	4.43	3.84	2.18	24.83

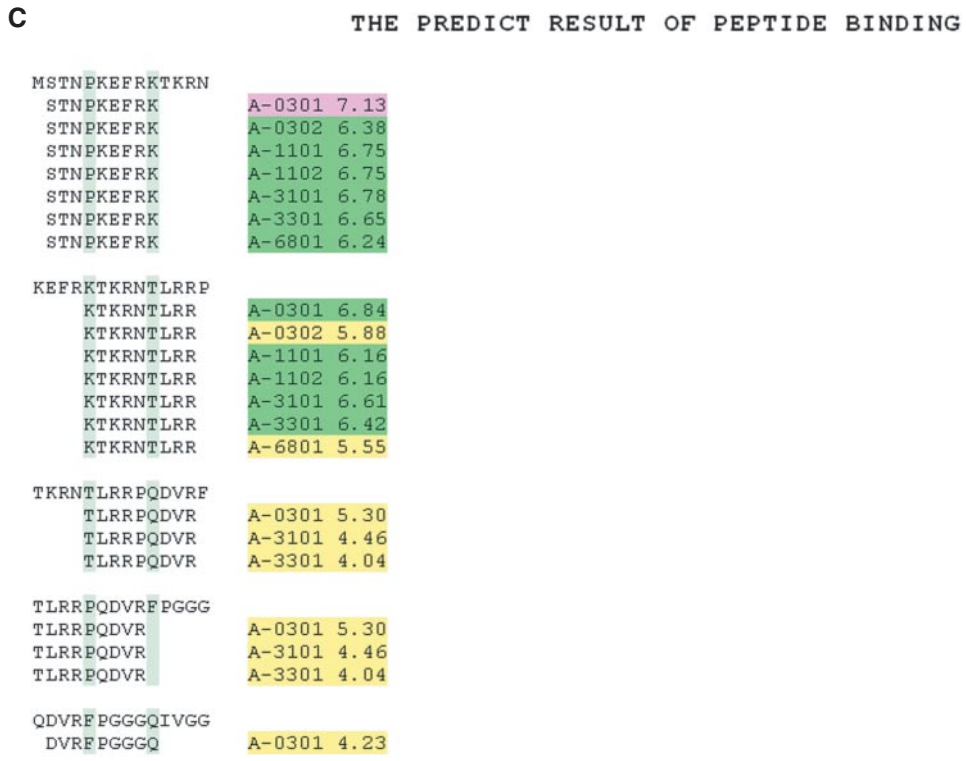


Figure 2. An example of the output pages of MULTIPRED when input is a list of peptides. The input protein peptides are from hepatitis C virus, the prediction method used is ANN and the HLA supertype of interest is HLA-A3. (**A**) The main result page. As can be seen here, the input sequence is truncated into overlapping 9mers for the prediction of binding scores to multiple HLA-A3 variants, *0301, *0302, *1101, *1102, *3101, *3301 and *6801. (**B**) Input peptides displayed in the descending order of binding scores. (**C**) Alignment view of the predicted 9mer binders.

prediction result table, 'Average' was calculated as the average of the 'Sum' within a 15mer window (seven consecutive 9mers make a 15mer window).

When users select the input sequence as 'a list of peptide sequences', the input sequences separated by carriage returns or line breaks are treated as different peptides. All overlapping 9mers in each peptide are submitted for prediction. In the result tables, predicted binding scores are represented by the highest individual binding score of each input peptide. The predicted binding scores of individual 9mers in each peptide in the list are data not shown (Figure 2A). To display the input peptides in the order of their binding scores, the user can use the function 'Sort the Result'. In the result page (Figure 2B), the input peptides are listed in descending order of their binding scores. To display the predicted 9mer binders from each input peptide, the user can use the function 'Alignment View'. In the result page (Figure 2C), the 9mers with binding scores $\geqslant 4$ are aligned with the input peptides. The predicted 9mer binders are displayed with the names of the HLA alleles, which produced binding scores above the selected threshold.

If the user has 9mer peptides with known binding affinities to proteins belonging to HLA-A2, -A3 or -DR supertypes and wants to build his own prediction models, the user can use the 'Model build' function in MULTIPRED. Only 9mer peptides can be used as training data. The users have the option to use their data only, or combine their data with the existing MULTIPRED data and build the model on the server. Currently, users can expect to train an HMM model within 1 min while training of ANN models may take up to 50 min (depending on the size of the training dataset)—there are actually four ANNs trained in the background. The ANN models trained by the same dataset are usually slightly different because the initial weights of networks are assigned randomly (14). To make the trained models more stable, the training is repeated four times, and four sets of weights are trained—the predictions are the averages of these four predictions. When the model building request is submitted, an intermediate page (Figure 3) will be displayed providing the result URL can be bookmarked for later model retrieval.

If the user has 9mer peptides with known binding affinities and would like to evaluate the prediction accuracy of a model with these peptides, the user can use the 'accuracy evaluation' function of MULTIPRED. The system predicts the binding affinities of the input 9mers and calculates A_{ROC} of the predictions. For each supertype, there are two built-in prediction models available in MULTIPRED. Predictions can also be performed by user-built models. Therefore, MULTIPRED may produce several sets of predictions for the same sequence. The comparison of predictions helps identify the most promising peptides picked up as predicted binders by multiple models. The comparison is facilitated by the 'Consensus predictions' function. The user needs to input the individual Result IDs (up to three) to the system. The Result IDs must be predictions of the same protein and to the same HLA supertype, if the Result IDs belong to predictions on different

Figure 3. When the model building request is submitted, an intermediate page will be displayed providing the result URL that can be bookmarked for later model retrieval.

proteins, an error message will be displayed. The user can select the analysis of top 5 or 10% of the predicted binders. In the output table, top 5 or 10% predictions are displayed in the descending order of their binding scores. The peptides selected by multiple models are highlighted in blue or red.

DISCUSSION

Several web-based systems have been developed and widely used for the prediction of MHC binders, such as SYFPEITHI (17), BIMAS (18), SMM (19), MHCPred (20), RANKPEP (21), TEPITOPE (22), NetMHC (23) and SVMHC (24). Although MULTIPRED is similar to them in its overall goal of predicting MHC-binding peptides, there are significant differences in both functionality and methodology. SYFPEITHI uses binding motifs. BIMAS, MHCpred, RANKPEP and TEPITOPE use quantitative matrices, and SMM is based on an improved matrix-based algorithm called stabilized matrix method. SVMHC uses support vector machines (SVMs) and NetMHC uses ANNs. Each of these methods uses one prediction model per MHC proteins, making them difficult to maintain and assess accuracy. TEPITOPE allows prediction of peptides to many different Class II proteins (using multiple prediction models), but it is not available through the Web. MULTIPRED predicts peptide binding to multiple HLA proteins with one model per HLA supertype. It can also identify promiscuous peptides and T-cell epitope hotspots. Since HLA proteins are highly polymorphic, promiscuous peptides that bind more than one HLA protein are prime targets for vaccine and immunotherapy development because they are relevant to higher proportions of the human population. T-cell epitope hotspots are highly promising regions as targets of T-cell immune responses, which are of interest for experimental validation as potential vaccine targets. In addition, MULTIPRED provides several functions which are not available in other prediction systems, such as model building by user function, accuracy evaluation function and consensus prediction function. The pathway from epitopes to vaccine development is lengthy and cost-intensive, involving exhaustive experiments. The main utility of MULTIPRED is in the selection of key antigenic regions to minimize the number of experiments required for mapping of promiscuous T-cell epitopes and T-cell epitope hotspots.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR online.

ACKNOWLEDGEMENTS

Authors thank Seng Hong Seah and Olivo Miotto for their valuable suggestions. This project has been funded in part with the USA Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Grant No. 5 U19 AI56541 and Contract No. HHSN2662-00400085C. Funding to pay the Open Access publication charges for this article was provided by the Institute for Infocomm Research, Singapore.

Conflict of interest statement. None declared.

REFERENCES

1. Pamer,E. and Cresswell,P. (1998) Mechanisms of MHC class I—restricted antigen processing. *Annu. Rev. Immunol.*, **16**, 323–358.
2. Villadangos,J.A., Bryant,R.A., Deussing,J. and Driessen,C. (1999) Proteases involved in MHC class II antigen presentation. *Immunol. Rev.*, **172**, 109–120.
3. Yewdell,J.W. and Bennink,J.R. (2001) Cut and trim: generating MHC class I peptide ligands. *Curr. Opin. Immunol.*, **13**, 13–18.
4. Bryant,P. and Ploegh,H. (2004) Class II MHC peptide loading by the professionals. *Curr. Opin. Immunol.*, **16**, 96–102.
5. Zhong,W., Reche,P.A., Lai,C.C., Reinhold,B. and Reinherz,E.L. (2003) Genome-wide characterization of a viral cytotoxic T lymphocyte epitope repertoire. *J. Biol. Chem.*, **278**, 45135–45144.
6. Marsh,S.G.E. (2004) Nomenclature for factors of the HLA system, update September 2003. *Tissue Antigens*, **63**, 190–191.
7. Sette,A. and Sidney,J. (1999) Nine major HLA class I supertypes account for the vast preponderance of HLA-A and -B polymorphism. *Immunogenetics*, **50**, 201–212.
8. Lund,O., Nielsen,M., Kesmir,C., Petersen,A.G., Lundegaard,C., Worning,P., Sylvester-Hvid,C., Lamberth,K., Røder,G., Justesen,S. *et al.* (2004) Definition of supertypes for HLA molecules using clustering of specificity matrices. *Immunogenetics*, **55**, 797–810.
9. Brusic,V., Bajic,V.B. and Petrovsky,N. (2004) Computational methods for prediction of T-cell epitopes—a framework for modelling, testing, and applications. *Methods*, **34**, 436–443.
10. Rammensee,H.G., Falk,K. and Rotzschke,O. (1993) Peptides naturally presented by MHC class I molecules. *Annu. Rev. Immunol.*, **11**, 213–244.
11. Stern,L.J., Brown,J.H., Jardetzky,T.S., Gorga,J.C., Urban,R.G., Strominger,J.L. and Wiley,D.C. (1994) Crystal structure of the human class II MHC protein HLA-DR1 complexed with an influenza virus peptide. *Nature*, **368**, 215–221.
12. Brusic,V., Rudy,G. and Harrison,L.C. (1994) MHCPEP, a database of MHC-binding peptides. *Nucleic Acids Res.*, **22**, 3663–3665.
13. Brusic,V., Petrovsky,N., Zhang,G.L. and Bajic,V.B. (2002) Prediction of promiscuous peptides that bind HLA class I molecules. *Immunol. Cell Biol.*, **80**, 280–285.
14. Zhang,G.L., Khan,A.M., Srinivasan,K.N., August,J.T. and Brusic,V. (2005) Neural models for predicting viral vaccine targets. *J. Bioinform. Comput. Biol.* (in press).
15. Srinivasan,K.N., Zhang,G.L., Khan,A.M., August,J.T. and Brusic,V. (2004) Predictions of Class I T-cell epitopes: evidence of presence of immunological hotspots inside antigens. *Bioinformatics*, **20** (Suppl. 1), i297–i302.

16. Kast,W.M., Brandt,R.M., Sidney,J., Drijfhout,J.W., Kubo,R.T., Grey,H.M., Melief,C.J. and Sette,A. (1994) Role of HLA-A motifs in identification of potential CTL epitopes in human papillomavirus type 16 E6 and E7 proteins. *J. Immunol.*, **152**, 3904–3912.
17. Rammensee,H.G., Bachmann,J., Emmerich,N.P., Bachor,O.A. and Stevanovic,S. (1999) SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics*, **50**, 213–219.
18. Parker,K.C., Bednarek,M.A. and Coligan,J.E. (1994) Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. *J. Immunol.*, **152**, 163–175.
19. Peters,B., Tong,W., Sidney,J., Sette,A. and Weng,Z. (2003) Examining the independent binding assumption for binding of peptide epitopes to MHC-I molecules. *Bioinformatics*, **19**, 1765–1772.
20. Hattotuwagama,C.K., Guan,P., Doytchinova,I.A., Zygouri,C. and Flower,D.R. (2004) Quantitative online prediction of peptide binding to the major histocompatibility complex. *J. Mol. Graph. Model*, **22**, 195–207.
21. Reche,P.A., Glutting,J.P. and Reinherz,E.L. (2002) Prediction of MHC class I binding peptides using profile motifs. *Hum. Immunol.*, **63**, 701–709.
22. Bian,H. and Hammer,J. (2004) Discovery of promiscuous HLA-II-restricted T cell epitopes with TEPITOPE. *Methods*, **34**, 468–475.
23. Buus,S., Lauemoller,S.L., Worning,P., Kesmir,C., Frimurer,T., Corbet,S., Fomsgaard,A., Hilden,J., Holm,A. and Brunak,S. (2003) Sensitive quantitative predictions of peptide-MHC binding by a 'Query by Committee' artificial neural network approach. *Tissue Antigens*, **62**, 378–384.
24. Donnes,P. and Elofsson,A. (2002) Prediction of MHC class I binding peptides, using SVMHC. *BMC Bioinformatics*, **3**, 25–38.