MetaMHCpan, a meta approach for pan-specific MHC peptide binding prediction

Yichang Xu[1,2], Cheng Luo[1,2], Hiroshi Mamitsuka[3], Shanfeng Zhu*[1,2]

[1]School of Computer Science, Fudan University, Shanghai 200433, China.

[2]Shanghai Key Lab of Intelligent Information Processing, Fudan University, Shanghai 200433, China.

[3]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Kyoto 611-0011, Japan.

Email: Yichang Xu - 12210240041@fudan.edu.cn;
Cheng Luo - 12210240095@fudan.edu.cn;
Hiroshi Mamitsuka -mami@kuicr.kyoto-u.ac.jp;
Shanfeng Zhu- zhusf@fudan.edu.cn;

*Corresponding author

MetaMHCpan, a meta approach for pan-specific MHC peptide binding prediction

Yichang Xu[1,2], Cheng Luo[1,2], Hiroshi Mamitsuka[3], Shanfeng Zhu*[1,2]

[1]School of Computer Science, Fudan University, Shanghai 200433, China.

[2]Shanghai Key Lab of Intelligent Information Processing, Fudan University, Shanghai 200433, China.

[3]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Kyoto 611-0011, Japan.

Abstract

Recent computational approaches in bioinformatics can achieve high performance, by which they can be a powerful support for performing real biological experiments, making biologists pay more attention to bioinformatics than before. In immunology, predicting peptides which can bind to MHC alleles is an important task, being tackled by many computational approaches. However this situation causes a serious problem for immunologists to select the appropriate method to be used in bioinformatics. To overcome this problem, we develop an emsemble prediction-based web server, which we call MetaMHCpan, consisting of two parts: MetaMHCIpan and MetaMHCIIpan, for predicting peptides which can bind MHC-I and MHC-II, respectively. MetaMHCIpan and MetaMHCIIpan use two (MHC2SKpan and LApan) and four (TEPITOPEpan, MHC2SKpan, LApan and MHC2MIL) existing predictors, respectively. MetaMHCpan is available at http://datamining-iip.fudan.edu.cn/MetaMHCpan/index.php/pages/view/info.

Keywords: MetaMHCpan, MHC-I, MHC-II, binding peptides, TEPITOPEpan, MHC2SKpan, MHC2MIL

1. Introduction

Major Histocompatibility Complex (MHC) and Human Leukocyte Antigen (HLA), a large family of genes in most vertebrates, plays important roles in adaptive immune response. An important function of MHC molecules is to bind peptide fragments derived from pathogens and to display the peptides on the cell surface to be recognized by the counterpart T cells [1]. Biochemical validation of peptides binding to MHC molecules is expensive and time consuming; while computational approaches are much more efficient, being recognized as useful, and allow to provide only a small number of top candidates (peptides) for further experimental verification.

Recent advances of immunoinformatics allow developing many computational methods for predicting peptides which can bind MHC molecules. These computational methods can be divided into two groups: allele-specific and pan-specific methods. Allele-specific methods train models by using binding data from an allele, and the model can be applied to predict peptides binding to the allele only. In this case if the number of binders for an allele is limited, the trained model for the allele is likely to fail to give a good predictive performance. To overcome this problem, the idea of pan-specific methods is to use data from multiple alleles as input and attempt to predict binders of not only the input alleles but also other alleles. In particular, this setting must be useful for predicting binders for alleles with very few or even no known binders [2, 3].

Currently several pan-specific methods have been proposed, which invites a problem of what methods are most reliable and should be used. To overcome this issue, we develop a web server, MetaMHCpan, an ensemble predictor using existing pan-specific methods as component predictors. MetaMHCpan

consists of MetaMHCIpan and MetaMHCIIpan, which predict peptides to bind to MHC-I and MHC-II, respectively. MetaMHCIpan uses twp pan- specific methods, MHC2SKpan [4] and LApan [5] for components, while MetaMHCIIpan uses three pan-specific methods: TEPITOPEpan [6], MHC2SKpan and LApan, and an allele-specific method: MHC2MIL [7] for components. Technically MetaMHCpan can achieve a higher predictive performance than component predictors, allowing MetaMHCpan to be current cutting-edge software on predicting peptide binders of a variety of MHC alleles.

2. Materials

The training set for MHC-I is Peters' dataset [8]. We use 35 HLA alleles and 6 H-2 alleles as our training alleles. Among these alleles, there are a total of 43312 peptides, and 12362 of them are binders. The training set for MHC-II is the dataset used by NetMHCIIpan-3.0 [9]. There are 24 DR alleles, 5 DP alleles, 6 DQ alleles and 2 H-2 alleles in this dataset with totally 52062 peptides, 20451 of which are binders.

3. Methods

MetaMHCIpan consists of two pan-specific methods: MHC2SKpan and LApan. MetaMHCIIpan consists of three pan-specific methods: TEPITOPEpan, MHC2SKpan and LApan, and one allele-specific method: MHC2MIL.

TEPITOPEpan is a position specific score matrix (PSSM) based method developed by extrapolating from the binding specifies of HLA-DR molecules characterized by TEPITOPE to those uncharacterized [6, 10]. The method can be divided into three steps: first, generating pseudo sequences of MHC binding pockets; then, computing the pocket similarity and weight between alleles; finally, computing PSSM. The predicted scores by TEPITOPEpan are not binding

affinities.

MHC2SKpan is a kernel based method. The string kernel MHC2SK (MHC-II String Kernel) used by MHC2SKpan measures the similarities among peptides with variable lengths [4]. We use support vector regression (SVR) as the predictor. The kernel for SVR is a product of an allele kernel and a peptide kernel (MHC2SK). The predicted scores by MHC2SKpan are binding affinities.

LApan is a method that we newly develop by extending the local alignment kernel (LA) [5] to a pan-specific method. The difference between LApan and MHC2SKpan is that the peptide kernel in LApan is LA kernel instead of MHC2SK. The predicted scores by LApan are binding affinities.

MHC2MIL is a multiple instance learning (MIL) based method by considering peptide flanking region and residue positions [7]. It is an allele-specific method and now we provide 35 alleles for prediction. Different from common supervised methods, MHC2MIL uses ‘bag’ instead of ‘instance’ to construct the learning unit. Each bag is mapped into a feature for SVR model with radial basis function (RBF) kernel. The predicted scores by MHC2MIL are binding affinities.

We also offer an integrated method AvgTanh [11, 12] to combine each selected method. TEPITOPEpan is a PSSM method, MHC2SKpan designs a new string kernel and MHC2MIL is a MIL based method. Since these methods are of different techniques, they are complement to each other and can get better results after integration. AvgTanh is an ensemble approach that the predicted score by each predictor of a test peptide will be converted into a Z-score first and then normalized by the tanh function. The final score will be the average of all

normalized scores.

3.1 MetaMHCIpan

MetaMHCIpan is for MHC-I peptide binding prediction. The input interface is shown in Figure 1.

1. Choose method. The default method is MHC2SKpan. LApan is another choice. At least one of the two methods should be chosen. AvgTanh is a complement if you want (See Note 1).
2. Choose input format. The default input format is FASTA Format. PEPTIDE Format is another choice (See Note 2).
3. Enter protein sequence(s). According to the data format chosen in step 2, enter proper peptides. If FASTA Format is chosen, please enter a long sequence with necessary information. If PEPTIDE Format is chosen, please enter several peptides line by line. The maximum number of peptides that can be accepted by the server is 500. You can upload a file instead of entering in the text area.
4. Select peptide length. If FASTA Format is chosen in step 2, please select the peptide length from 9-mer to 11-mer. The default value is 9-mer. The sequence in step 3 will be cut according the peptide length you select. If PEPTIDE Format is chosen, no peptide length should be chosen since the peptides are already entered in a proper length in step 3.
5. Select species and loci. For Human, HLA-A or HLA-B can be a choice. For Mouse, H-2 is a choice.
6. Select allele. According to the species and loci chosen in step 5, different alleles will be the candidates. For HLA-A, 19 alleles from HLA-A0101 to HLA-A6901 can be selected. For HLA-B, 16 alleles from HLA-B0702 to HLA-B5801 can be selected. For H-2, 6 alleles from H-2-Db to H-2-Ld can be selected.
7. Input your MHC-I sequence. You can input your MHC-I sequence in FASTA format if previous species and alleles do not meet your demand. You can upload a file instead of entering in the text area.
8. Choose output interface. The output can be displayed on the webpage or in a text format. The default output interface is webpage. It is easy for you to read

while the text format is more convenient for a computer program to analyze.

9. Click submit button. Click the submit button at the bottom of the page and your task will be in processing. You can reset all by clicking the reset bottom aside. If your task takes a little bit long time, you can input your email address and the result will be sent to you by email. The results are IC50 in nm. Peptides with IC50 less than 500nm can be deemed as a binder. Rank will be displayed aside if “show rank” button is clicked (See Note 3).

3.2 MetaMHCIIpan

MetaMHCIIpan is for MHC-II peptide binding prediction. The input interface is shown in Figure 2.

1. Choose method. The default methods are MHC2SKpan and LApan. MHC2MIL or TEPITOPEpan can be other choices. At least one of the four methods should be chosen. AvgTanh is a complement if you want (See Note 1).
2. Choose input format. The format is the same as MetaMHCIpan. (See Note 2).
3. Enter protein sequence(s). It is the same as MetaMHCIpan.
4. Select peptide length. If FASTA Format is chosen in step 2, please select the peptide length from 9-mer to 25-mer. The default setting is 15-mer. The sequence in step 3 will be cut according the peptide length you select.
5. Select species and loci. For Human, one of HLA-DP, HLA-DQ, HLA-DRB1, HLA-DRB3, HLA-DRB4 and HLA-DRB5 can be a choice. For Mouse, H-2 is a choice.
6. Select allele. Different alleles are provided as a list to be chosen according to the species and loci decided in step 5.
7. Input your MHC-II sequence. You can input your MHC-II sequence in the FASTA format if previous species and alleles do not meet your demand. Two MHC chains should be input. They are alpha chain and beta chain. You can upload a file instead of entering in the text area.
8. Chose output interface. It is the same as MetaMHCIpan.
9. Click submit button. Different from other methods, the results of TEPITOPEpan is not IC50 nm. They are the scores predicted by TEPITOPEpan. The rank

may help you to judge the binding ability of peptides (See Note 3).

4. Notes

1. In MetaMHCIpan, the score of AvgTanh is the average Tanh score of MHC2SKpan and LApan. In MetaMHCIIpan, the score of AvgTanh is the average Tanh score of MHC2SKpan, LApan and MHC2MIL.

2. The button, “show an example”, on the input page can give you some examples.

3. Figure 3 and Figure 4 are output examples on webpage. The first section “Prediction Finished” shows the chosen allele and the time cost. The second section “Prediction Results” displays the results by the MetaMHCpan. You can choose “Plain Format” by clicking the red line “Show this Table in Plain Format” on the top of this section. The results are displayed by a table with pagination. You can choose how many entries you want in a page. The first column of table is the peptide number. The second column is the corresponding peptide. This is followed by the columns with results by different methods you have chosen. You can sort the results as you like by clicking the column names at the top of the table. You can also click the button “show rank” or “hidden rank” to show or hide the rank of predictions. “Search” function is used to search key words such as peptide sequence.

Acknowledgements

This work has been partially supported by National Natural Science Foundation of China (Grant Nos: 61170097), and Scientific Research Starting Foundation for Returned Overseas Chinese Scholars, Ministry of Education, China.

References

1. Janeway J CA, Travers P, Walport M, et al (2001) Immunobiology: The Immune System in Health and Disease. New York: Garland Science Publishing 5 edition
2. Zhang L, Udaka K, Mamitsuka H, Zhu S (2012) Toward more accurate pan-specific MHC-peptide binding prediction: a review of current methods and tools. Briefings in bioinformatics 13(3):350–364
3. Zhu S, Udaka K, Sidney J, Sette A, Aoki-Kinoshita KF, Mamitsuka H (2006) Improving MHC binding peptide prediction by incorporating binding data of auxiliary MHC molecules. Bioinformatics 22(13):1648–1655
4. Guo L, Luo C, Zhu S (2013) MHC2SKpan: a novel kernel based approach for pan-specific MHC class II peptide binding prediction. BMC Genomics 14(Suppl 5):S11
5. Salomon J, Flower DR (2006) Predicting Class II MHC-Peptide binding: a kernel based approach using similarity scores. BMC bioinformatics 7:501
6. Zhang L, Chen Y, Wong HS, Zhou S, Mamitsuka H, Zhu S (2012) TEPITOPEpan: extending TEPITOPE for peptide binding prediction covering over 700 HLA-DR molecules. PLoS One 7(2):e30483
7. Xu Y, Luo C, Qian M, Huang X, Zhu S (2014) MHC2MIL: a novel multiple instance learning based method for MHC II peptide binding prediction by considering peptide flanking region and residue positions. BMC Genomics 15(Suppl 9):S9
8. Peters B, Bui HH, Frankild S, Nielsen M, Lundegaard C, Kostem E, Basch D, Lamberth K, Harndahl M, Fleri W, et al (2006) A community resource benchmarking predictions of peptide binding to MHC-I molecules. PLoS computational biology 2(6):e65
9. Karosiene E, Rasmussen M, Blicher T, Lund O, Buus S, Nielsen M (2013) NetMHCIIpan-3.0, a common pan-specific MHC class II prediction method including all three human MHC class II isotypes, HLA-DR, HLA-DP and HLA-DQ. Immunogenetics 65(10):711–724
10. Sturniolo T, Bono E, Ding J, Raddrizzani L, Tuereci O, Sahin U, Braxenthaler M, Gallazzi F, Protti MP, Sinigaglia F, et al (1999) Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. Nature biotechnology 17:555–561
11. Hu X, Zhou W, Udaka K, Mamitsuka H, Zhu S (2010) MetaMHC: a meta approach to predict peptides binding to MHC molecules. Nucleic acids research gkq407
12. Hu X, Mamitsuka H, Zhu S (2011) Ensemble approaches for improving HLA Class I-peptide binding prediction. Journal of immunological methods 374:47–

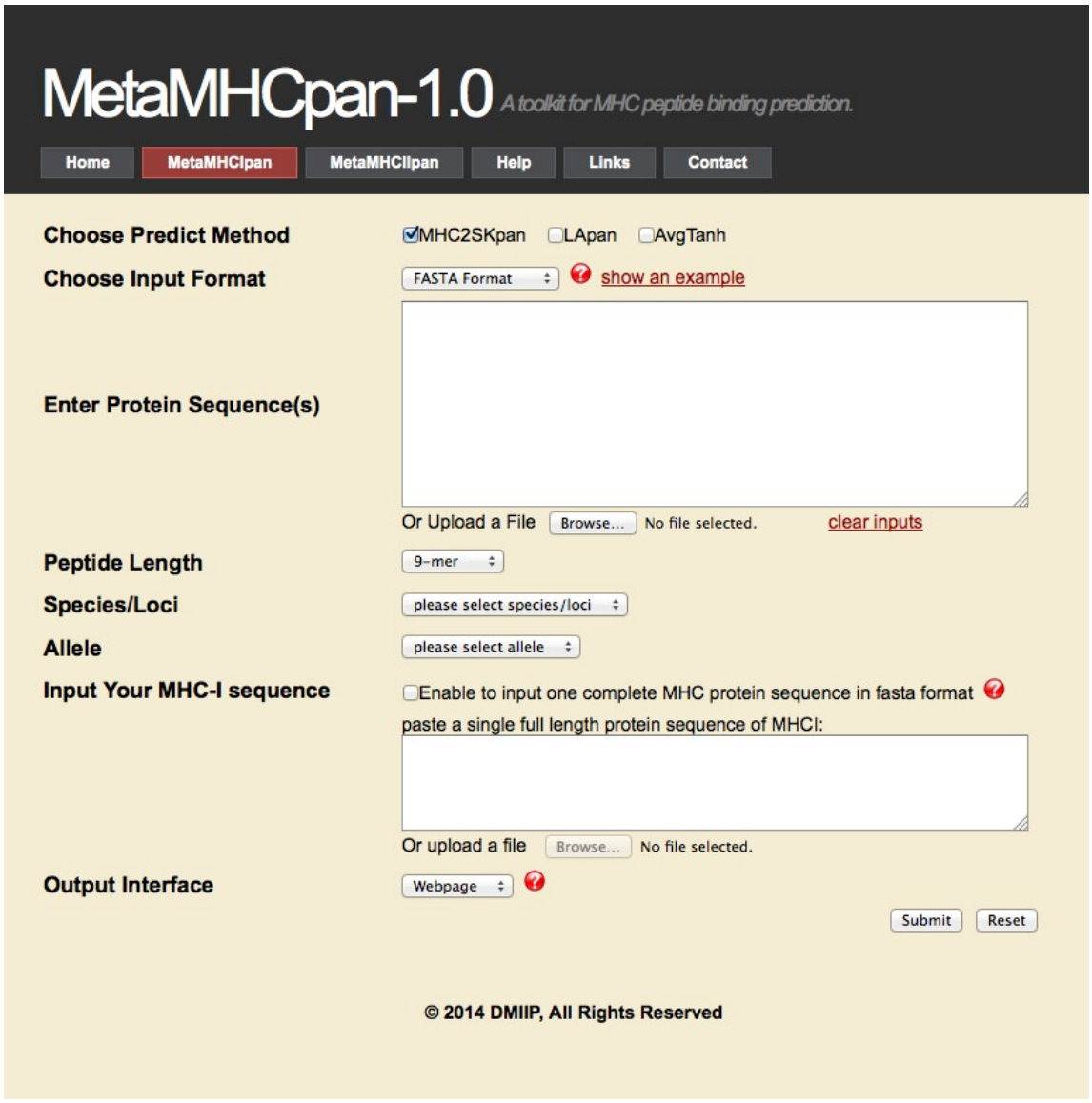

Figure 1: Input interface for MetaMHCIpan

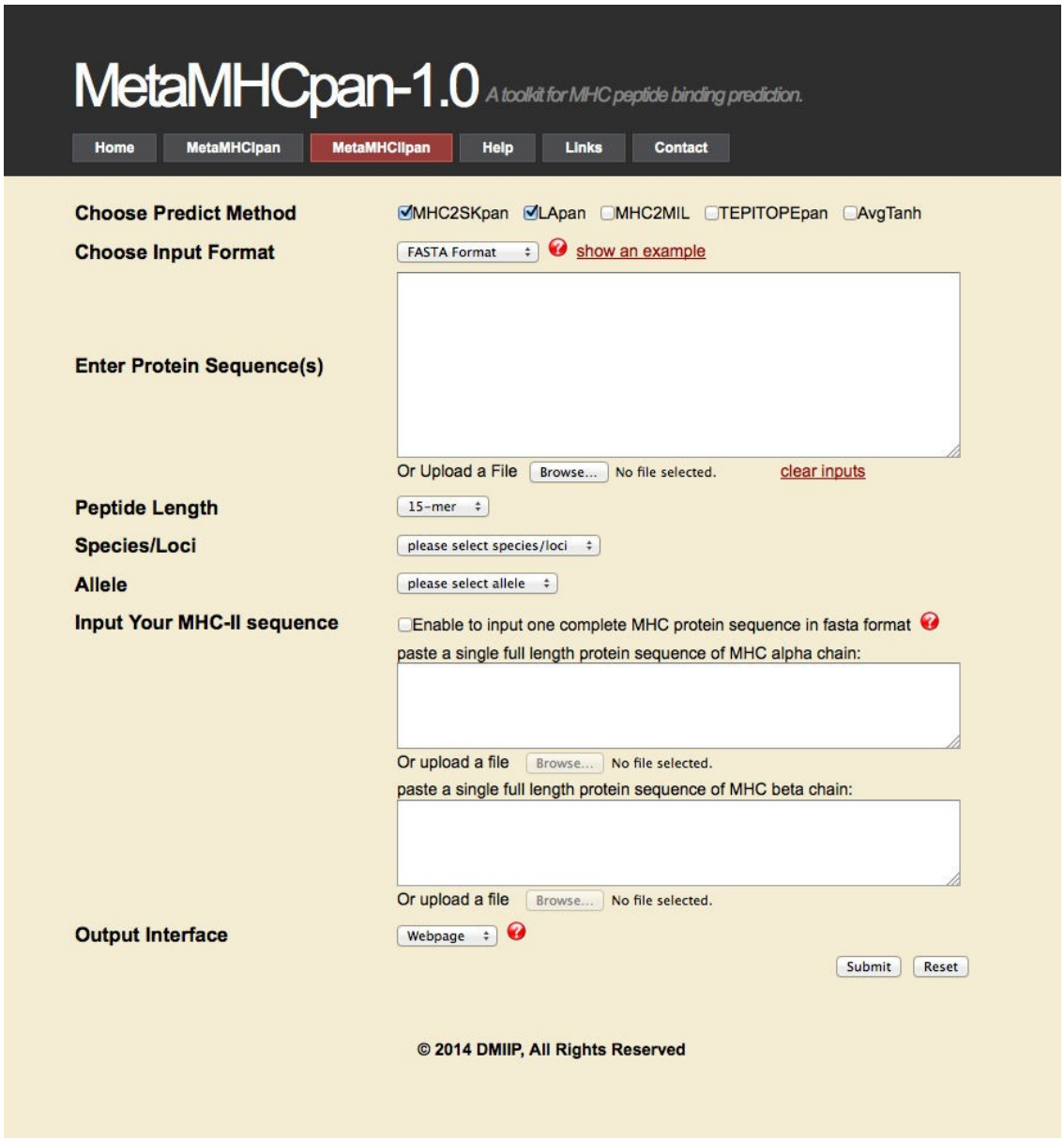

Figure 2: Input interface for MetaMHCIIpa

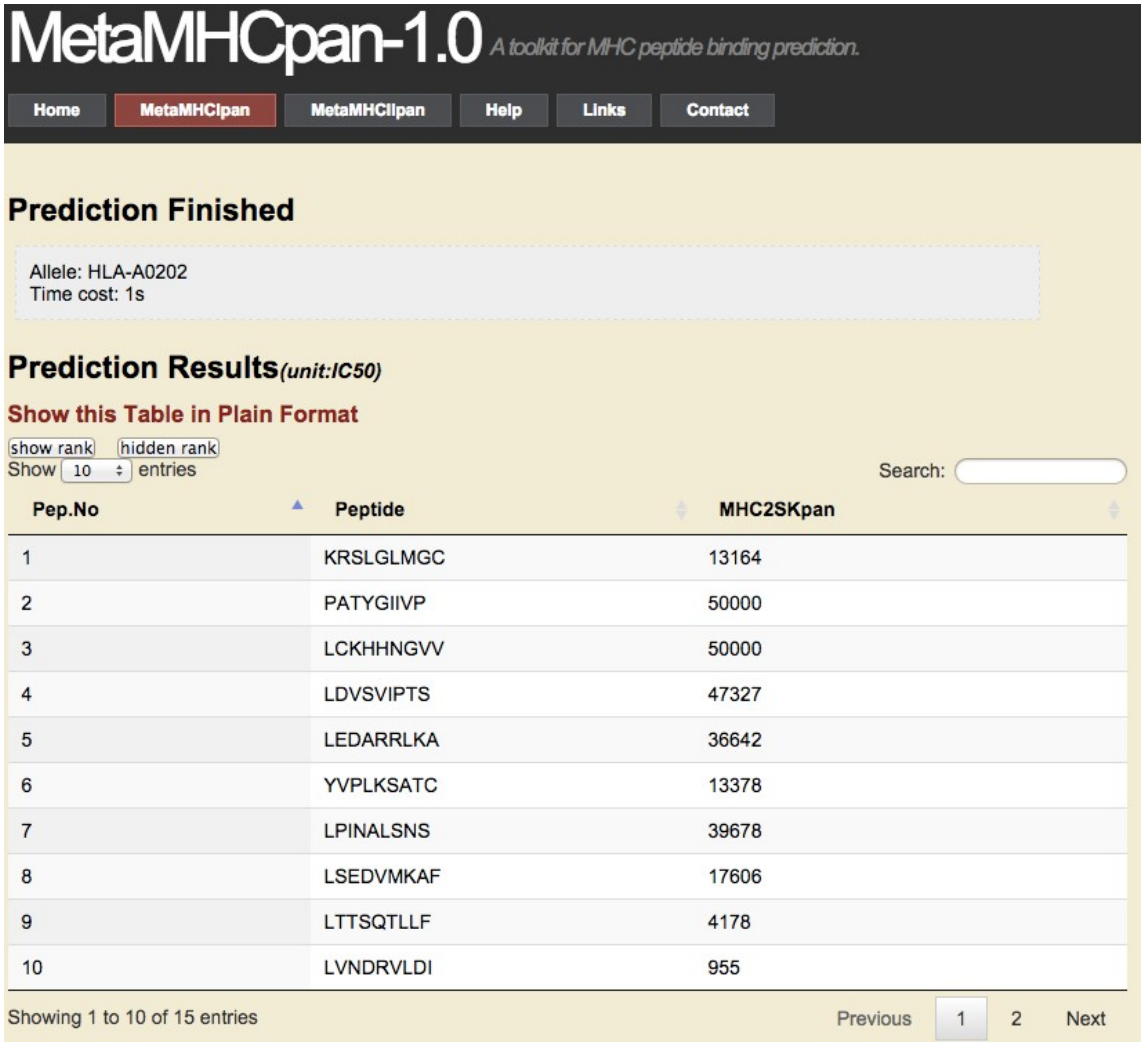

Figure 3: output example for MetaMHCIIpan

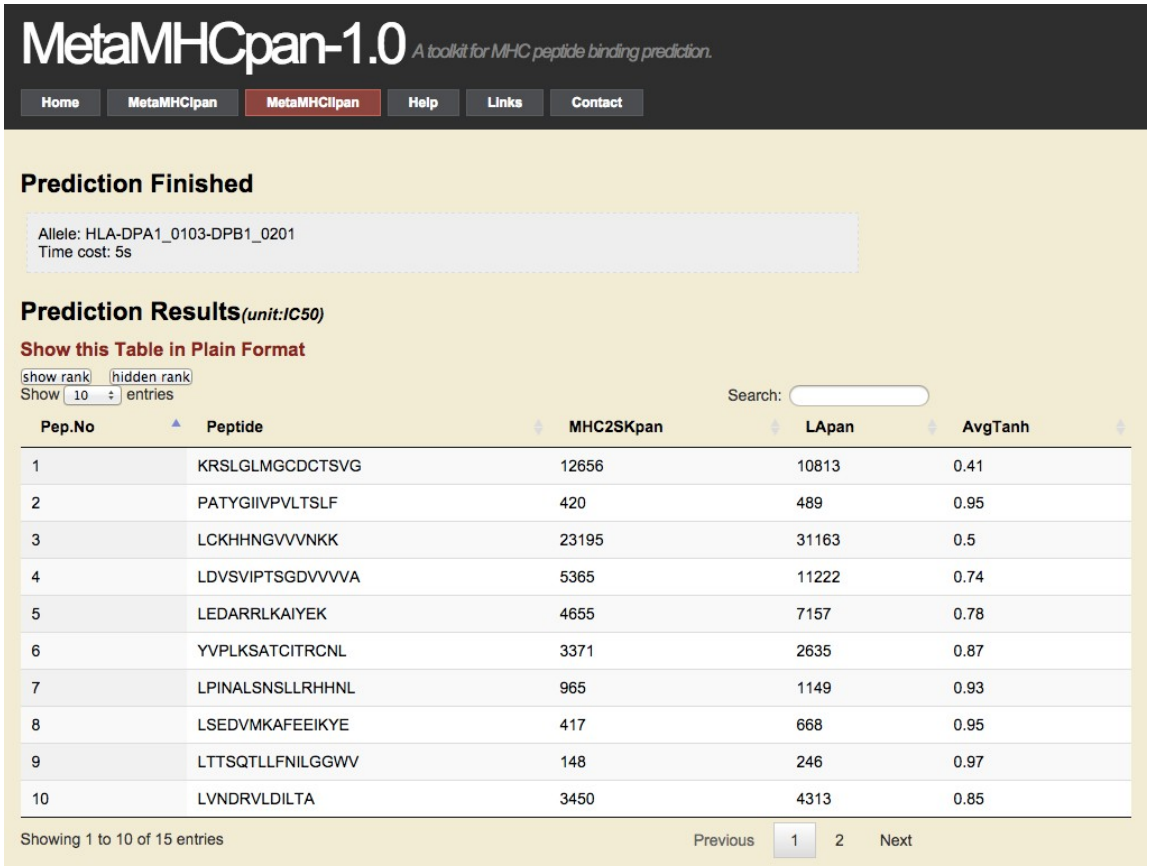

Figure 4: output example for MetaMHCIIpan