HyperBinding

A Peptide-Class I Major Histocompatibility Complex (MHC) Binding Prediction Tool

Jingyi Xie, Jinrong Ma, Bowei Zhang, Xiaofeng Xiang
UW DIRECT project

Project Overview

Hyperbinding is a MHC ligand prediction python package using machine learning approach to generate high-confidence peptides by considering the presentation possibilities of peptides with MHC molecules. MHC-peptide binding prediction plays an important role in cancer immunoengineering, T cell therapy and vaccine design.

A simplified immune recognition process involves the binding of a short peptide with antigen presenting cells, where the peptide comes from cancer neoantigen or virus antigen. To be specific, T cells recognize fragments of protein antigens that have been partly degraded inside the antigen-presenting cell. The peptide fragments are then carried to the surface of the presenting cell on special molecules called MHC proteins, which present the fragments to T cells. For the Class I MHC, the binding groove is optimal for peptides with 8 to 12 amino acids. Higher binding affinity between MHC and the peptide will allow the MHC-peptide complex to be recognized by T cells and thus elicit an immune cascade in eliminating the pathogen.

The current release was trained using only data from HLA-A02:01, which is one of the most abundant alleles in Class I human MHC molecules. Currently, the training and testing support peptides ranging from 8 to 12 amino acids in length, but the model can be re-trained to support more alleles and wider ranges of peptide length. The further application of HyperBinding can be extended into different Class I MHC alleles, and broad categories of disease-related peptides.

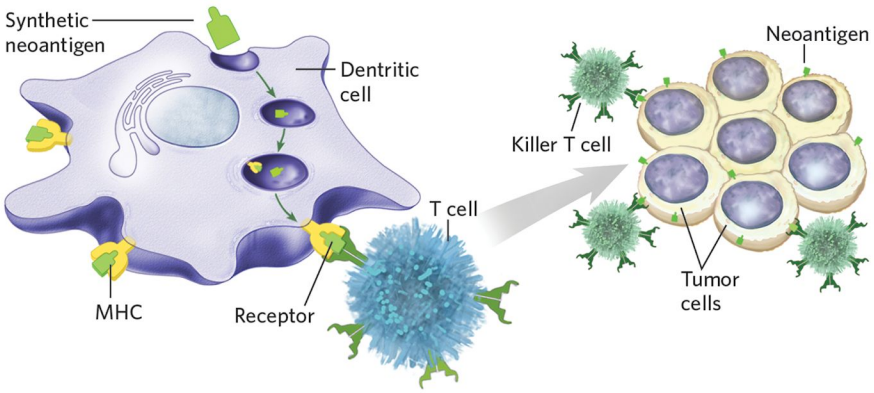


Fig. 1 Antigen presenting and T cell recognition in the immune system.

Potential Users

The prediction of peptide-MHC binding affinity has rapidly accelerated the development of vaccines and adoptive T-cell therapies targeting viruses. The primary users of HyperBinding are researchers who want to get a list of peptides which are potent Class I MHC A0201 binder.

- Primary User Example:

An outbreak of respiratory disease caused by a novel coronavirus which has now been detected in more than 100 locations internationally, including in the United States. The virus has been named “SARS-CoV-2” and the disease it causes has been named “coronavirus disease 2019” (abbreviated “COVID-19”). William is an Bioengineer working on T cell engineering and immunotherapy. His lab is supporting the research on COVID-19 by designing high-throughput capture reagents to select and isolate T cells with potent immunity in recognizing and eliminating the virus. Then the isolated T cells will be sequenced and amplified to be transferred back to patients to boost the immune system against the virus.

To design the high-throughput capture reagents, William needs a list of peptides that are 8 mers from the Surface Spike Glycoprotein (S protein) of SARS-CoV-2 virus. The patient is HLA-A02:01 positive (indicating the type of MHC molecule) so these peptides should also have higher binding affinity with HLA-A02:01.

Without being overwhelmed by complicated installation and coding, HyperBindng is an accessible tool to help him in generating the list of potent binders of HLA-A02:01. William has the protein sequence of SARS-CoV-2 virus, which is made up of 1272 amino acids. Now he can follow the instructions to slice the protein sequence into fragments with desirable length and predicate their binding ability to HLA-A02:01.

- Other Potential User Example:

Tumor-specific neoantigens have attracted much attention since they can be used as biomarkers to predict therapeutic effects of immune checkpoint blockade therapy and as potential targets for cancer immunotherapy. Next-generation sequencing data from tumor and normal DNA are aligned and compared to the human reference genome and then to each other to identify tumor-specific alterations. These variants are then evaluated for their resultant changes to the amino acid sequences of the encoded proteins.

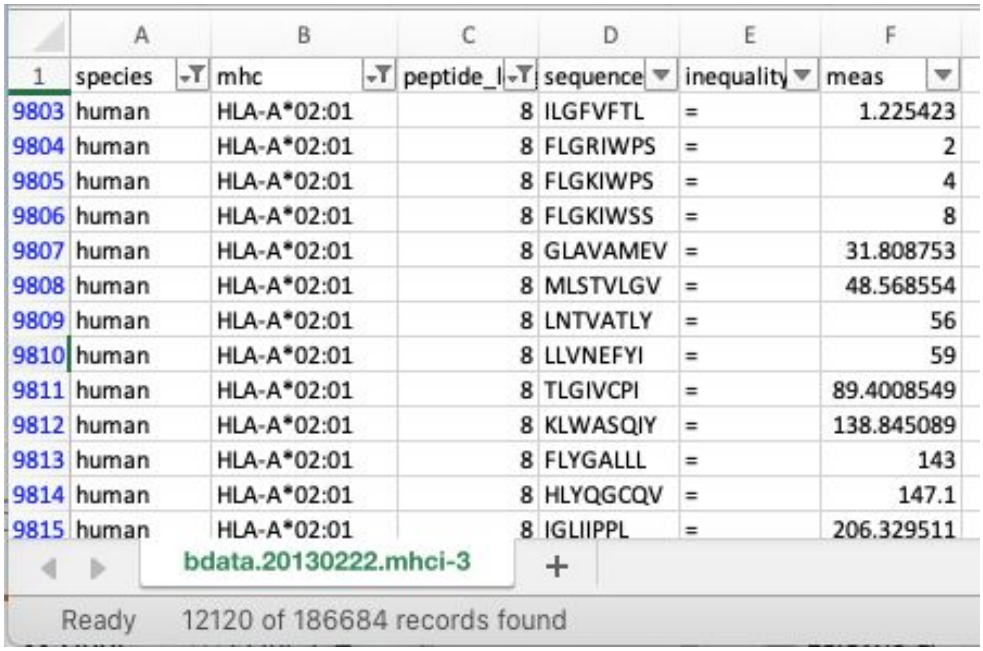
Rick is working on cancer immunotherapy for prostate cancer. For a recent clinical trial, the patient tumor sample was sequenced and analyzed for somatic mutation data. Based on the alignment result, there are 6,961 potential neoantigens generated by somatic mutations of the tumor sample. To study the T cell response against these neoantigens, Rick needs to narrow the scope of potential neoantigens by selecting the neoantigens with higher binding affinity with the patient's MHC type (HLA-A02:01). The

selected sequences can be evaluated by HyperBinding that predict the binding of the neoantigens to HLA-A02:01 that would present them on the surface of cells. The prediction result would facilitate the personalized neoantigen-based cancer immunotherapy.

Data Description

The data we are using for model training and testing are from IEDB Analysis Resource and are available for public (http://tools.iedb.org/main/).

The dataset was compiled from three sources: the IEDB, the Sette lab, and the Buus lab. If a peptide/allele combination had more than 1 measurement among the three sources, its geometric mean was taken. The data was a compressed text file containing binding data for different species including human, Chimpanzee, Gorilla, etc. We extract data for humans and HLA-A02:01 for MHC type for our project. As we mentioned, the binding groove of a Class I MHC molecule allows binding of peptides with 8-12 amino acids. So we generate a new data set of peptides with lengths of 8 to 12, and their binding affinity. In this data set, 12120 peptides with their sequences and binding affinity are included.



	A	B	C	D	E	F
1	species	mhc	peptide_l	sequence	inequality	meas
9803	human	HLA-A*02:01	8	ILGFVFTL	=	1.225423
9804	human	HLA-A*02:01	8	FLGRIWPS	=	2
9805	human	HLA-A*02:01	8	FLGKIWPS	=	4
9806	human	HLA-A*02:01	8	FLGKIWSS	=	8
9807	human	HLA-A*02:01	8	GLAVAMEV	=	31.808753
9808	human	HLA-A*02:01	8	MLSTVLGV	=	48.568554
9809	human	HLA-A*02:01	8	LNTVATLY	=	56
9810	human	HLA-A*02:01	8	LLVNEFYI	=	59
9811	human	HLA-A*02:01	8	TLGIVCPI	=	89.4008549
9812	human	HLA-A*02:01	8	KLWASQIY	=	138.845089
9813	human	HLA-A*02:01	8	FLYGALLL	=	143
9814	human	HLA-A*02:01	8	HLYQGCQV	=	147.1
9815	human	HLA-A*02:01	8	IGLIIPPL	=	206.329511

bdata.20130222.mhci-3

Ready 12120 of 186684 records found

Fig 2. Original data for analysis

Design Strategy

HyperBinding is a package of files and algorithms to generate a feature map for each peptide sequence to interpret its biochemical properties, including peptide sequence, length, hydropathy index, polarity and size. The feature map will be used as the input of a preliminary convolution neural network for machine learning. Then the binding affinity of the peptide-MHC complex will be predicted based on the training model.

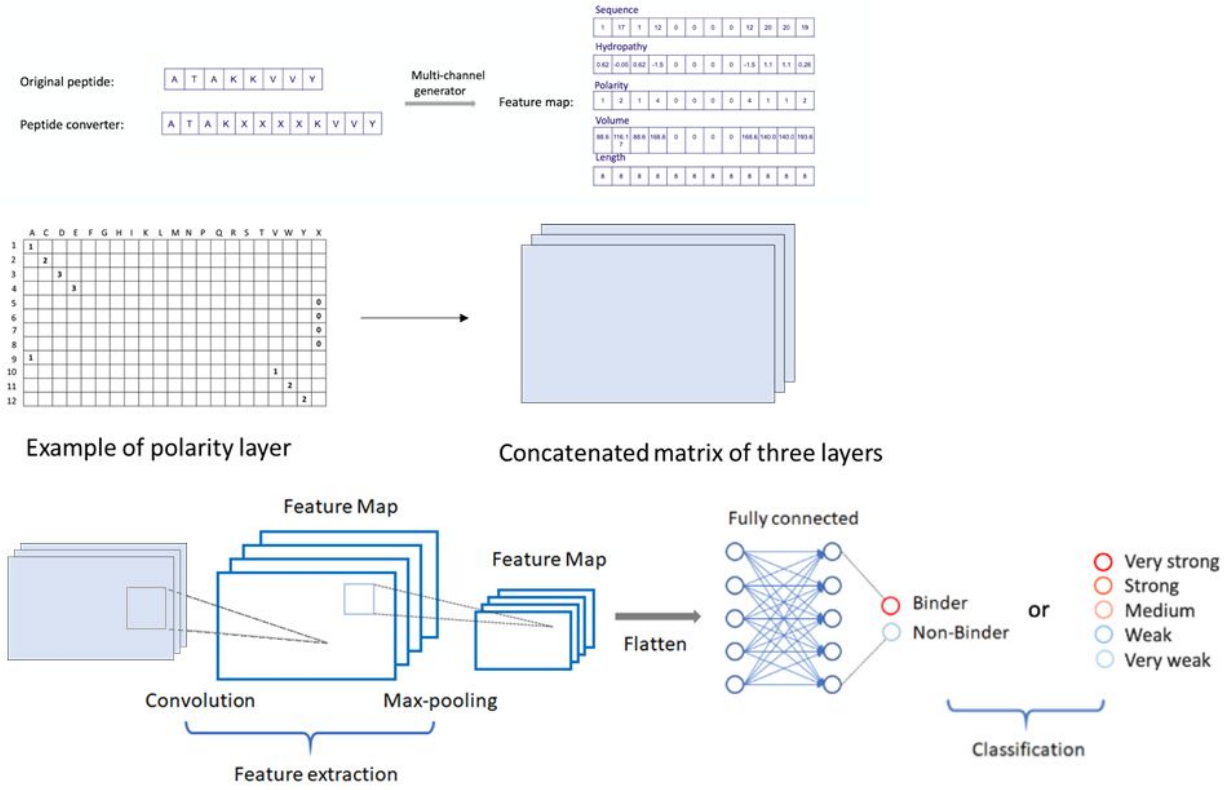


Fig 3. Design strategy and CNN

For the Class I MHC, the binding groove is optimal for peptides with 8 to 12 amino acids. Peptides presented by a Class I MHC molecule generally assume a central bulged conformation. Based on this theory, we proposed a novel method that can convert the 8–12 mer peptide to uniform size of 12. Since the peptide residues on both sides are much more important than the other locations, we try to ensure that the new ‘amino acid’ (X) is only inserted into the middle position of the peptide.

The main function for Hyperbinding is predicting the binding affinity of peptide toward HLA-A02:01. The chemical properties of peptides have been reported to strongly affect the binding affinity. Since we consider that the order of the sequence, hydropathy index, polarity and the length of the peptide could affect the binding affinity and the properties of these amino acids are key factors for their binding to MHC, we extracted these information from each peptide sequence.

For the polarity index, we divided 21 amino acids into five classes. According to the polarity of R group or the trend of interaction with water at physiological pH (approaching pH 7.0), they can be divided into non-polarity, polarity without charge, positive charge (alkalinity) and negative charge (acidity). X's class is zero. Similarly, each amino acid will have a hydropathy index , and a volume index. Then we can convert a 12 mer peptide sequence into a 5x12 input matrix or 21x12x3 input matrix for CNN analysis. Quantitative values of amino acid properties are shown in Table 1 and are used in generating input matrix.

Amino acid	Abbreviations		Molecular mass (Da)	Number of atoms	Volume (A^3) [1]	Hydropathy index [2]
Alanine	Ala	A	89	13	88.6	1.8
Arginine	Arg	R	174	26	173.4	-4.5
Asparagine	Asn	N	132	17	114.1	-3.5
Aspartic acid	Asp	D	133	16	111.1	-3.5
Asparagine or Aspartic acid	Asx	B				
Cysteine	Cys	C	121	14	108.5	2.5
Glutamine	Gln	Q	146	20	143.8	-3.5
Glutamic Acid	Glu	E	147	19	138.4	-3.5
Glutamine or Glutamic acid	Glx	Z				
Glycine	Gly	G	75	10	60.1	-0.4
Histidine	His	H	155	20	153.2	-3.2
Isoleucine	Ile	I	131	22	166.7	4.5
Leucine	Leu	L	131	22	166.7	3.8
Lysine	Lys	K	146	24	168.6	-3.9
Methionine	Met	M	149	20	162.9	1.9
Phenylalanine	Phe	F	165	23	189.9	2.8
Proline	Pro	P	115	17	112.7	-1.6
Serine	Ser	S	105	14	89.0	-0.8
Threonine	Thr	T	119	17	116.1	-0.7
Tryptophan	Trp	W	204	27	227.8	-0.9
Tyrosine	Tyr	Y	181	24	193.6	-1.3
Valine	Val	V	117	19	140.0	4.2

Class	Label	Amino acids
NONE	0	X
Polarity without charge	1	A, G, I, L, F, P, V
Non-polarity	2	N, C, Q, S, T, W, Y, M
Negative charge (acidity)	3	D, E
Positive charge (alkalinity)	4	R, H, K

Table 1. Quantitative informations of different amino acids

- For the 5x12 input matrix, the first row stands for the index of peptide sequence, second row stands for quantitative value of hydrophoty, third row for polarity, fourth row for volume and fifth row for peptide length.
- For the 21x12x3 input matrix, there are three channels. First channel stands for hydropathy layer, second for volume layer and third for polarity layer. In each layer, there are 21 rows and 12 columns. Rows stand for different amino

acids(totally 21) and columns stand for 12 peptide sequence locations. In a single peptide, we first figure out the location of each amino acid in the 21x12 (single layer) and assign its corresponding property value. The rest of blanks assign a value of zero. After doing these for each amino acid in peptide, a 21x12x3 matrix is generated and then is ready to be applied in CNN.

Primary Application

Model fitting and Prediction:

1. Using this tool, a simple input of peptide sequence can be used for five category binding prediction by running

HyperBinding/examples/prediction/main_multi_prediction.ipynb

Welcome to HyperBinding. We can help you to predict your peptide binding affinity towards MHC. Our method works for peptide with a length short than 12. Please follow below instruction to get your result.

In this method:
very strong binder has kd<=50nM,
strong binder have 50nM<kd<=500nM,
medium binder have 500nM<kd<=1000nM,
weak binder have 1000nM<kd<=20000nM,
very weak binder have kd>20000nM

```
: print('Please enter your protein sequence')
  sequence = input()
```

```
Please enter your protein sequence
ACFRLSGK
```

And you can get a prediction of how strong the peptide will bind to the MHC molecule:

```
your sequence is weak binder
```

2. Using this tool, a simple input of peptide sequence can be used for a binary binding prediction by running

HyperBinding/examples/prediction/main_binary_prediciton.ipynb

Welcome to HyperBinding. We can help you to predict your peptide binding affinity towards MHC. Our method works for peptide with a length short than 12. Please follow below instruction to get your result.

In this method, strong binder has kd<= 500nM and weak binder have kd>500nM

```
print('Please enter your protein sequence')
sequence = input()
```

```
Please enter your protein sequence
ACFRLSGK
```

Then you can get a binary prediction whether the peptide is a strong binder or not.

```
your sequence does not have strong binding affinity to MHC
```

3. Using this tool, a long protein sequence can be sliced into short peptides for binding affinity prediction by running

HyperBinding/examples/prediction/main_binding_predictor.ipynb

Welcome to HyperBinding. We can help you slice your protein into fragments with desirable length and predicate their binding affinity to HLA-A-02.

```
print('Please enter your protein sequence')
sequence = input()
```

```
Please enter your protein sequence
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFDNPVLPFNDGVYFASTEKS
NIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFCNDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKN
IDGYFKIYSKHTPINLVRDLPQGFSALEPLVDLPIGINITRFQTLLALHRSYLTPGDSSSGWTAGAAAYYVGYLQPRTFLLKYNENGTITDAVDCALD
PLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCF
TNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYF
PLQSYGFQPTNGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDIT
PCSFGGVSVITPGTNTSNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNSPRRARS
VASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQ
VKQIYKTPPIKDFGGFNFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTI
TSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVVNQNAQALNTLVKQLSSNFGAISSVLNDI
LSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTA
PAICHDGKAHFPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASV
VNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
```

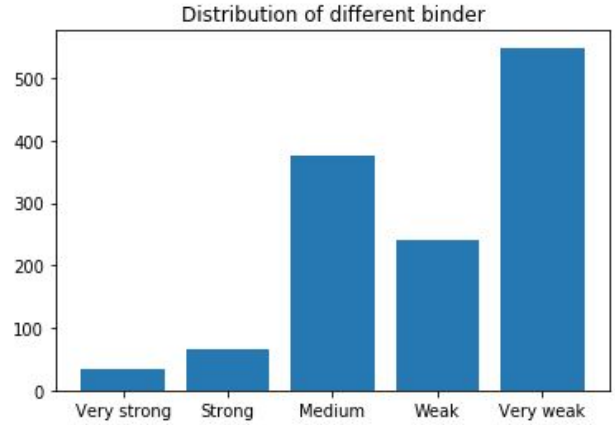
```
print('Please enter the length of your sliced fragments')
length = int(input())
```

```
Please enter the length of your sliced fragments
8
```

```
Please enter which class of binding you want to see:
for very strong binder, kd<=30nM, please enter 0
for strong binder, 50nM<kd<=500nM, please enter 1
for medium binder, 500nM<kd<=10000nM, please enter 2
for weak binder, 10000nM<kd<=20000nM, please enter 3
for very weak binder, kd>20000nM, please enter 4
0
```

Then you will get a list of peptides with assigned length and binding affinity with the MHC molecule.

Your protein will be sliced into 8 -mers, total 1266
Below are your result summary:



The fragments of Very strong binders are shown below
MFVFLVLL
FLVLLPLV
VLLPLVSS
SFTRGVYY
KVFRSSVL
DLFLPFFS
FFSNVTWF
RFDNPVLP
FQFCNDPF
DPFLGVYY
CTFEYVSQ
FLMDLEGK
GFSALEPL
YLTPGDSS
YYVGYLQP
YVGYLQPR
YLQPRTFL
YQAGSTPC
ELLHAPAT
FNFNGLTG
CSFGGVSV
YQDVNCTE
YTMSLGAE
FTISVTTE
FSQILPDP
GLTVLPPL
YTSALLAG
WTFGAGAA
KLQDVVNQ
HLMSFPQS
MSFPQSAP
FLHVTYVP
GTHWFVTQ
YIKWPWYI
FIAGLIAI