HyperBinding

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

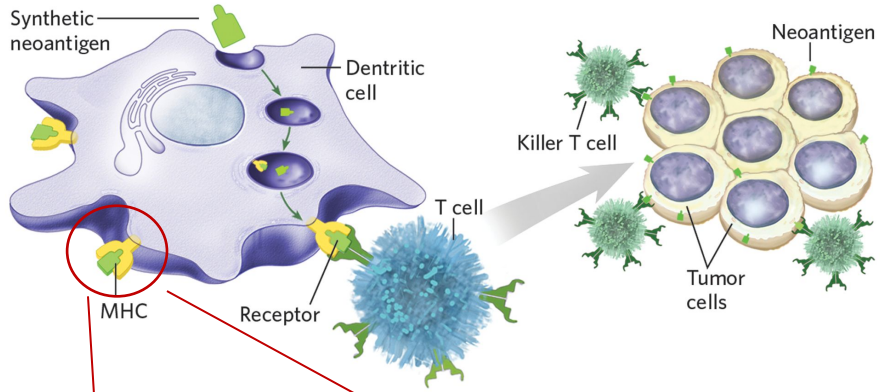
Jingyi Xie, Jinrong Ma, Bowei Zhang, Xiaofeng Xiang

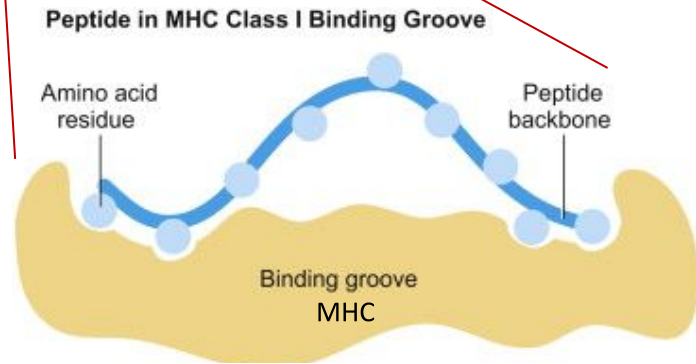
UNIVERSITY of WASHINGTON

Biological Background for Peptide-Major Histocompatibility Complex (MHC)

Antigen-presenting cells display peptide-MHC complex on cell surface to trigger antigen-specific T cell response against tumor cells or virus infected cells.

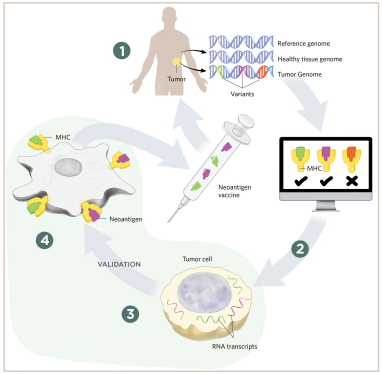
Peptides (8-12 amino acids) presented by a MHC-I molecules (represented as ribbon cartoons) generally assume a central bulged conformation.





https://www.the-scientist.com/features/personalized-cancer-vaccines-in-clinical-trials-66075

Peptide-Class I MHC Binding Prediction Tool and Our Dataset



Biomedical Applications:
Peptide-based vaccine
Adoptive T cell therapy
…

Our aim:
Use machine learning methods to predict binding affinity

Quick screening of the binding affinity between MHC proteins and their peptide ligands.

	A	B	C	D	E	F
1	species	mhc	peptide_length	sequence	inequality	meas
2	human	HLA-A*02:01	8	ILGFVFTL	=	1.225423
3	human	HLA-A*02:01	8	FLGRIWPS	=	2
4	human	HLA-A*02:01	8	FLGKIWPS	=	4
5	human	HLA-A*02:01	8	FLGKIWSS	=	8
6	human	HLA-A*02:01	8	GLAVAMEV	=	31.808753
7	human	HLA-A*02:01	8	MLSTVLGV	=	48.568554
8	human	HLA-A*02:01	8	LNTVATLY	=	56
9	human	HLA-A*02:01	8	LLVNEFYI	=	59
10	human	HLA-A*02:01	8	TLGIVCPI	=	89.4008549

Dataset:
http://tools.iedb.org/main/
The dataset was compiled from three sources: the IEDB, the Sette lab, and the Buus lab.

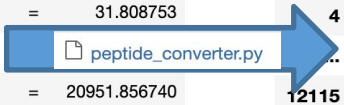
Class I MHC: Human HLA-A*02:01 (One abundant serotype)
Peptide length: 8-12 mer
Peptide counts: 12120

Convert all sequence to a uniform length

```python
import pandas as pd
df = pd.read_csv('HLA-A*0201.csv')
df
```

	species	mhc	peptide_length	sequence	inequality	meas
0	human	HLA-A*02:01	8	ILGFVFTL	=	1.225423
1	human	HLA-A*02:01	8	FLGRIWPS	=	2.000000
2	human	HLA-A*02:01	8	FLGKIWPS	=	4.000000
3	human	HLA-A*02:01	8	FLGKIWSS	=	8.000000
4	human	HLA-A*02:01	8	GLAVAMEV	=	31.808753
...	...	...	...	...	...	...
12115	human	HLA-A*02:01	12	SLVWAPLILAYF	=	20951.856740
12116	human	HLA-A*02:01	12	LVGKLNWASQIY	=	23904.872640
12117	human	HLA-A*02:01	12	DPHGPVQLSYYD	>	69444.444440
12118	human	HLA-A*02:01	12	LYDSQGLPEELP	>	69444.444440
12119	human	HLA-A*02:01	12	TNIRQAGVQYSR	>	178571.428600

12120 rows × 6 columns




```python
import pandas as pd
df = pd.read_csv('Filled-HLA-A*0201.csv', index_col=0)
df
```

	species	mhc	peptide_length	sequence	inequality	meas
0	human	HLA-A*02:01	8	ILGFXXXXVFTL	=	1.225423
1	human	HLA-A*02:01	8	FLGRXXXXIWPS	=	2.000000
2	human	HLA-A*02:01	8	FLGKXXXXIWPS	=	4.000000
3	human	HLA-A*02:01	8	FLGKXXXXIWSS	=	8.000000
4	human	HLA-A*02:01	8	GLAVXXXXAMEV	=	31.808753
...	...	...	...	...	...	...
12115	human	HLA-A*02:01	12	SLVWAPLILAYF	=	20951.856740
12116	human	HLA-A*02:01	12	LVGKLNWASQIY	=	23904.872640
12117	human	HLA-A*02:01	12	DPHGPVQLSYYD	>	69444.444440
12118	human	HLA-A*02:01	12	LYDSQGLPEELP	>	69444.444440
12119	human	HLA-A*02:01	12	TNIRQAGVQYSR	>	178571.428600

12120 rows × 6 columns

Label peptides according to equilibrium constant

```
1 import pandas as pd
2 df = pd.read_csv('Filled-HLA-A*0201.csv', index_col=0)
3 df
```

	species	mhc	peptide_length	sequence	inequality	meas
0	human	HLA-A*02:01	8	ILGFXXXXVFTL	=	1.225423
1	human	HLA-A*02:01	8	FLGRXXXXIWPS	=	2.000000
2	human	HLA-A*02:01	8	FLGKXXXXIWPS	=	4.000000
3	human	HLA-A*02:01	8	FLGKXXXXIWSS	=	8.000000
4	human	HLA-A*02:01	8	GLAVXXXXAMEV	=	31.808753
...	...	...	...	...	...	...
12115	human	HLA-A*02:01	12	SLVWAPLILAYF	=	20951.856740
12116	human	HLA-A*02:01	12	LVGKLNWASQIY	=	23904.872640
12117	human	HLA-A*02:01	12	DPHGPVQLSYYD	>	69444.444440
12118	human	HLA-A*02:01	12	LYDSQGLPEELP	>	69444.444440
12119	human	HLA-A*02:01	12	TNIRQAGVQYSR	>	178571.428600

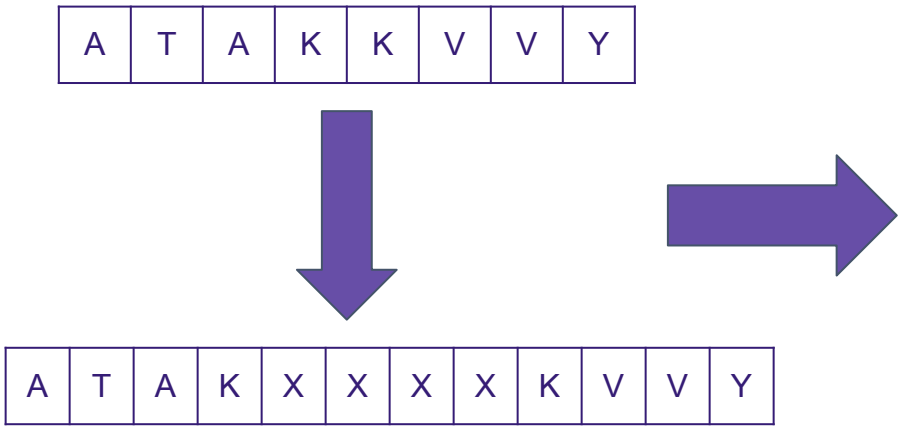
12120 rows × 6 columns


```
1 import pandas as pd
2 df = pd.read_csv('Labeled-Filled-HLA-A*0201.csv', index_col=0)
3 df
```

	species	mhc	peptide_length	sequence	inequality	meas	label
0	human	HLA-A*02:01	8	ILGFXXXXVFTL	=	1.225423	P
1	human	HLA-A*02:01	8	FLGRXXXXIWPS	=	2.000000	P
2	human	HLA-A*02:01	8	FLGKXXXXIWPS	=	4.000000	P
3	human	HLA-A*02:01	8	FLGKXXXXIWSS	=	8.000000	P
4	human	HLA-A*02:01	8	GLAVXXXXAMEV	=	31.808753	P
...	...	...	...	...	...	...	...
12112	human	HLA-A*02:01	12	TLVGLAIGLVLL	=	698.917985	N
12113	human	HLA-A*02:01	12	DILSGIFSNPHP	=	900.407007	N
12114	human	HLA-A*02:01	12	SDILSGIFSNPH	=	13131.280920	N
12117	human	HLA-A*02:01	12	DPHGPVQLSYYD	>	69444.444440	N
12118	human	HLA-A*02:01	12	LYDSQGLPEELP	>	69444.444440	N

8950 rows × 7 columns

Transform each peptide into a characteristic matrix



Feature (5 X 12)

Sequence	1	17	1	12	0	0	0	0	12	20	20	19
Hydropathy	0.62	-0.05	0.62	-1.5	0	0	0	0	-1.5	1.1	1.1	0.26
Polarity	1	2	1	4	0	0	0	0	4	1	1	2
Volume	88.6	116.17	88.6	168.6	0	0	0	0	168.6	140.0	140.0	193.6
Length	8	8	8	8	8	8	8	8	8	8	8	8

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

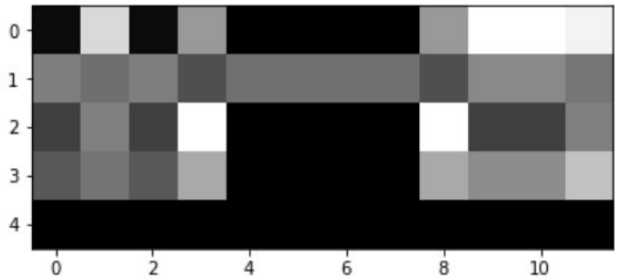
Transform characteristic matrix to feature grayscale images

1	17	1	12	0	0	0	0	12	20	20	19
0.62	-0.05	0.62	-1.5	0	0	0	0	-1.5	1.1	1.1	0.26
1	2	1	4	0	0	0	0	4	1	1	2
88.6	116.17	88.6	168.6	0	0	0	0	168.6	140.0	140.0	193.6
8	8	8	8	8	8	8	8	8	8	8	8

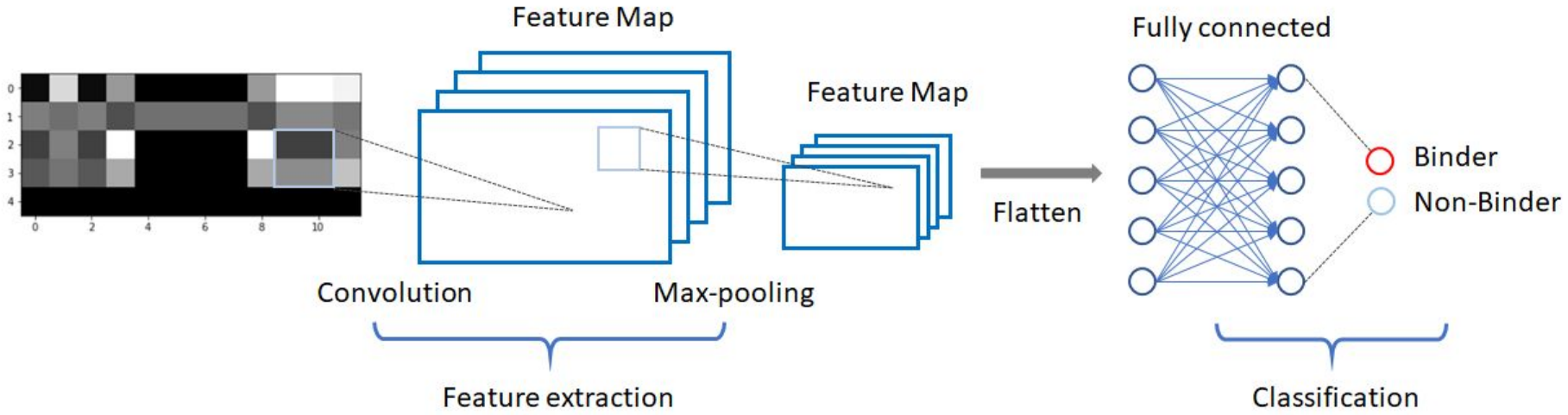
Normalize to (0,255)

123	217	13	153	0	0	0	0	153	255	255	243
127	112	127	79	113	113	113	113	138	138	138	119
64	128	64	255	0	0	0	0	255	64	64	128
88.6	116.17	88.6	168.6	0	0	0	0	168.6	140.0	140.0	193.6
0	0	0	0	0	0	0	0	0	0	0	0

Transform to grayscale img



Using Convolutional Neural Network (CNN) for machine learning



Package: Keras
https://github.com/keras-team/keras

Packages to used

- Statistics: Pandas

  Manipulating excel format database
  Show the distribution of database
  Already learned in class

- Machine learning: Kears

  Easy to used for image based machine learning
  Works well for previous project