

## The prediction problem for MLI competition:

Training data (peptides and their binding affinity) will be used to train a prediction system. An example of training data for 9-mer peptides is shown in Table 1.

### Peptide:

Defined as a string of nine characters over the alphabet consisting of 20 symbols (A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, Y, and W).

### Binding:

Peptide binding is defined as the strength of force that bonds the peptide and the HLA molecule. Low values of  $IC_{50}$  correspond to stronger binding and high values correspond to low binding. The scaled scores are mainly between 0 and 100, where values of 0 correspond to no binding and 100 represents binding of a known high-affinity probe. Some peptides that have ultra-high binding might have scores higher than 100. An extract from a representative example is shown in Table 1.

### Prediction system:

A classifier is trained using training data sets (e.g. those available at the DFRMLI site: <http://bio.dfci.harvard.edu/DFRMLI/>).

When presented with peptides whose binding affinity is not known, the classifiers typically produce scores for each peptide. These scores are used for prediction of binding. The best general source for peptide binding affinity is the IEDB database (Zhang *et al.*, 2008).

### Prediction task:

Predict binding affinity; the scaled score will be used as the experimental value.

### Assessment of accuracy:

Accuracy will be assessed using  $A_{ROC}$  (also known as the Area Under the Curve, or  $A_{UC}$ ). For details see study by Lin, *et al.* (2008).

**Table 1.** An example showing representative data points for the binding prediction task. Selected peptides and their binding affinity to HLA-B\*0801 are shown.

Molecule: B*0801			
Peptide	Length	$IC_{50}$ (nM)	Scaled Score
RPYGKFRAM	9	67.3	100
YLKKWLNSF	9	69.5	100
IPRRNVATL	9	71.0	100
YRYLRHGKL	9	104.5	99
ILLRKGHVF	9	107.2	99
...	...	...	...
EIKFNDITF	9	2639.7	82
IVAWTRTAT	9	2717.5	82
KLFIRQEEV	9	2848.3	81
IVAWTRTAT	9	2848.3	81
...	...	...	...
MTMRRRLF	9	5634.7	62
DIIRAHDPWF	9	9919.0	34
YTFCLNVK	9	10630.3	29
LPLIVDTAA	9	11150.0	26
RMLPKLAEF	9	11847.5	21
VMLLDIDYF	9	13305.0	11
ATFSRPGSL	9	13400.0	11
RRRQWASCM	9	13885.0	7
KSYEHQTPF	9	14415.0	4
DYAMHGTVF	9	14576.7	3
TMPELAWAV	9	15000.0	0
...	...	...	...
HSKKCCDEL	9	76632.9	0
HSKRCCDEL	9	77563.2	0
HSNIEEVAL	9	77777.8	0
VKKLWGHLP	9	185200.0	0

## Data sets

Several independent data sets have been made available for training, testing, and benchmarking. The data sets available at the DFRMLI site are shown in Table 2.

**Table 2.** Available data sets at DFRMLI. These data originate from Peters *et al.* (2006), and Nielsen *et al.* (2003). Benchmark data for B\*0801 and B\*1501 10mers are not available.

Molecule	Peptides (length)	Training set sizes	Benchmarked?	Aroc (area under the ROC curve)
A*0101	9	1157 + 447	-	
A*0101	10	56	-	
A*0201	9	3088 + 444	+	0.95
A*0201	10	1316	-	
A*0301	9	2092 + 331	+	0.94
A*0301	10	1082	-	
A*1101	9	1983 + 219	+	0.92
A*1101	10	1093	-	
A*2402	9	195 + 367	+	0.82
A*2402	10	78	-	
B*0702	9	1261 + 232	+	0.98
B*0702	10	205	-	
B*0801	9	708 + 119	+	0.95
B*1501	9	976 + 114	+	0.93

## References

Lin HH, Ray S, Tongchusak S, Reinherz EL, Brusci V. Evaluation of MHC class I peptide binding prediction servers: applications for vaccine research. *BMC Immunol.* **2008**;9:8.

<http://www.biomedcentral.com/1471-2172/9/8>

Nielsen M, Lundegaard C, Worning P, Lauemøller SL, Lamberth K, Buus S, Brunak S, Lund O. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations.

*Protein Sci.* **2003**;12(5):1007-17. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=12717023>

Peters B, Bui HH, Frankild S, Nielson M, Lundegaard C, Kostem E, Basch D, Lamberth K, Harndahl M, Fleri W, Wilson SS, Sidney J, Lund O, Buus S, Sette A. A community resource benchmarking predictions of peptide binding to MHC-I molecules. *PLoS Comput Biol.* **2006**;2(6):e65.

<http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.0020065>

Zhang Q, Wang P, Kim Y, Haste-Andersen P, Beaver J, Bourne PE, Bui HH, Buus S, Frankild S, Greenbaum J, Lund O, Lundegaard C, Nielsen M, Ponomarenko J, Sette A, Zhu Z, Peters B. Immune epitope database analysis resource (IEDB-AR). *Nucleic Acids Res.* **2008**; 36:W513-8.

[http://nar.oxfordjournals.org/cgi/content/full/36/suppl\\_2/W513](http://nar.oxfordjournals.org/cgi/content/full/36/suppl_2/W513)