Supplementary Files of

DeepSeqPan, a sequence-based deep convolutional neural network model for peptide-HLA binding affinity prediction

Zhonghao Liu[1], Yuxin Cui[1], Zheng Xiong[1], Alierza Nasiri[1], Ansi Zhang[2], Jianjun Hu[1,2*]

1 Department of Computer Science and Engineering, University of South Carolina, 29201 Columbia, SC, United States
2 School of Mechanical Engineering, Guizhou University, 50033 Guiyang, Guizhou, China

***Corresponding, jianjunh@cse.sc.edu**

1 Availability

All dataset, source code and trained models are included in our Github repository https://github.com/pcpLiu/DeepSeqPan.

2 Dataset

2.1 Training dataset

The training dataset is downloaded from http://tools.iedb.org/main/datasets/, binding_data_2013.zip.

2.2 Benchmark testing dataset

We downloaded all available weekly benchmark dataset from http://tools.iedb.org/auto_bench/mhci/weekly/ including from 2014-03-21 to 2018-01-26. Since there are quite a lot duplicates between training and benchmark dataset, we removed all duplicates in benchmark dataset. The filtered dataset can be found in Github repository.

3 5-fold Cross validation results

The 5-fold cross validation results measured for each allele are listed in the Following Table 1. In the table, some AUC scores are N/A because of that in training dataset, there are few alleles coming with 0 positive samples. So it is not applicable to calculate AUC scores. For alleles HLA-B*27:10 and HLA-A*02:04, the SRCC cannot be calculated since their number of samples ≤ 2 which is not a calculable situation for spearmanr function defined in Scipy (https://docs.scipy.org/doc/scipy-0.18.1/reference/generated/scipy.stats.mstats.spearmanr.html). For alleles HLA-A*02:10, the Spearmanr function gives nan result because all samples are labeled as same value which is also a non-calculable situation define in Scipy.

4 Compared with Kim's DCNN model

As show in Table 2, we listed AUC scores of Kim's DCNN model and our proposed DeepSeqPan.

5 DeepSeqPan network details

5.1 Layer configuration

FC *N*. A fully connected layer with N hidden units.
Dropout. We used 0.5 as dropout rate.
ConvBlock *N*. A ConvBlock N consists of 4 layers in order:

1. A 2D Convolutional layer with N filters of size 1×2;
2. A Batch normalization layer;
3. A 2D Convolutional layer with N filters of size 1×2;
4. A Max pooling with kernel size 1×2.

LCBlock *N*. A LCBlock N consists of 3 layers in order:

1. A 2D Locally Connected layer with N filters of size 1×2;
2. A Batch normalization layer;
3. A MLeakyReLU activation layer.

5.2 Parameters initialization and regularizations

For all trainable layers, we use uniform distribution in range $[-0.05, 0.05]$ to randomly generate initial values of all parameters. And we disabled bias parameters for all trainable layers. $l2$ regularizer is used for all parameter training and the control factor is 0.1.

Table 1. 5-fold cross validation measured for each allele

HLA	IC_{50}		Binary		HLA	IC_{50}		Binary	
	AUC	SRCC	AUC	SRCC		AUC	SRCC	AUC	SRCC
HLA-A*01:01	0.93	0.57	0.94	0.55	HLA-B*15:09	0.88	0.47	0.89	0.47
HLA-A*02:01	0.95	0.81	0.95	0.81	HLA-B*15:17	0.93	0.69	0.94	0.69
HLA-A*02:02	0.93	0.86	0.93	0.86	HLA-B*15:42	0.82	0.04	0.65	-0.01
HLA-A*02:03	0.95	0.82	0.95	0.82	HLA-B*18:01	0.90	0.52	0.89	0.48
HLA-A*02:04	N/A	N/A	N/A	N/A	HLA-B*27:01	0.00	-1.00	0.00	-1.00
HLA-A*02:05	0.99	0.84	0.98	0.81	HLA-B*27:02	N/A	-0.50	N/A	-0.50
HLA-A*02:06	0.91	0.80	0.91	0.79	HLA-B*27:03	N/A	0.03	N/A	-0.01
HLA-A*02:07	0.76	0.75	0.79	0.79	HLA-B*27:04	N/A	1.00	N/A	1.00
HLA-A*02:10	N/A	N/A	N/A	N/A	HLA-B*27:05	0.94	0.63	0.94	0.59
HLA-A*02:11	0.96	0.80	0.96	0.80	HLA-B*27:06	N/A	-1.00	N/A	-1.00
HLA-A*02:12	0.97	0.77	0.97	0.77	HLA-B*27:10	N/A	N/A	N/A	N/A
HLA-A*02:16	0.98	0.68	0.98	0.68	HLA-B*27:20	0.56	0.52	0.54	0.56
HLA-A*02:17	0.68	0.41	0.69	0.41	HLA-B*35:01	0.90	0.72	0.91	0.71
HLA-A*02:19	0.96	0.66	0.96	0.66	HLA-B*35:03	0.87	0.50	0.90	0.54
HLA-A*02:50	1.00	0.88	0.99	0.87	HLA-B*37:01	0.48	0.01	0.61	0.13
HLA-A*03:01	0.93	0.73	0.93	0.72	HLA-B*38:01	0.98	0.81	0.98	0.80
HLA-A*03:02	0.69	0.57	0.68	0.57	HLA-B*39:01	0.93	0.62	0.93	0.61
HLA-A*03:19	0.87	0.55	0.87	0.56	HLA-B*40:01	0.97	0.63	0.97	0.59
HLA-A*11:01	0.95	0.77	0.95	0.76	HLA-B*40:02	0.90	0.76	0.90	0.75
HLA-A*11:02	1.00	0.77	1.00	0.77	HLA-B*40:13	0.64	0.48	0.61	0.46
HLA-A*23:01	0.93	0.71	0.92	0.68	HLA-B*42:01	0.94	0.77	0.94	0.77
HLA-A*24:02	0.90	0.67	0.90	0.66	HLA-B*42:02	0.84	0.64	0.82	0.63
HLA-A*24:03	0.96	0.70	0.96	0.68	HLA-B*44:02	0.95	0.60	0.94	0.55
HLA-A*25:01	0.98	0.47	0.99	0.47	HLA-B*44:03	0.94	0.82	0.95	0.82
HLA-A*26:01	0.93	0.52	0.93	0.49	HLA-B*45:01	0.92	0.67	0.91	0.66
HLA-A*26:02	0.96	0.74	0.96	0.74	HLA-B*45:06	0.81	0.14	0.71	0.17
HLA-A*26:03	0.93	0.52	0.94	0.54	HLA-B*46:01	0.92	0.44	0.93	0.44
HLA-A*29:02	0.88	0.65	0.87	0.63	HLA-B*48:01	0.92	0.49	0.92	0.50
HLA-A*30:01	0.91	0.70	0.92	0.70	HLA-B*51:01	0.92	0.58	0.92	0.55
HLA-A*30:02	0.82	0.63	0.82	0.62	HLA-B*52:01	0.58	0.29	0.50	0.18
HLA-A*31:01	0.93	0.74	0.93	0.74	HLA-B*53:01	0.92	0.78	0.93	0.77
HLA-A*32:01	0.85	0.71	0.85	0.72	HLA-B*54:01	0.90	0.73	0.90	0.73
HLA-A*32:07	0.81	0.53	0.82	0.53	HLA-B*57:01	0.96	0.62	0.96	0.59
HLA-A*32:15	0.51	0.38	0.50	0.36	HLA-B*57:02	0.84	0.66	0.80	0.62
HLA-A*33:01	0.92	0.73	0.92	0.73	HLA-B*57:03	0.97	0.74	0.97	0.75
HLA-A*66:01	0.83	0.41	0.84	0.35	HLA-B*58:01	0.95	0.69	0.96	0.68
HLA-A*68:01	0.90	0.79	0.90	0.79	HLA-B*58:02	0.55	0.51	0.59	0.55
HLA-A*68:02	0.92	0.70	0.92	0.69	HLA-B*73:01	0.64	0.36	0.66	0.37
HLA-A*68:23	0.77	0.53	0.78	0.54	HLA-B*81:01	0.92	0.76	0.93	0.76
HLA-A*69:01	0.94	0.51	0.94	0.53	HLA-B*83:01	0.93	0.53	0.93	0.53
HLA-A*74:01	0.70	0.45	0.61	0.33	HLA-C*03:03	0.80	0.59	0.79	0.58
HLA-A*80:01	0.94	0.55	0.95	0.55	HLA-C*04:01	0.52	-0.09	0.58	-0.15
HLA-B*07:02	0.95	0.72	0.95	0.72	HLA-C*05:01	0.91	0.74	0.92	0.75
HLA-B*08:01	0.91	0.66	0.91	0.65	HLA-C*06:02	0.89	0.74	0.89	0.74
HLA-B*08:02	0.96	0.42	0.96	0.44	HLA-C*07:01	0.84	0.61	0.84	0.61
HLA-B*08:03	0.92	0.35	0.95	0.38	HLA-C*07:02	0.74	0.42	0.73	0.41
HLA-B*14:01	0.74	0.43	0.74	0.44	HLA-C*08:02	0.64	0.28	0.63	0.25
HLA-B*14:02	0.84	0.51	0.85	0.43	HLA-C*12:03	0.62	0.28	0.61	0.29
HLA-B*15:01	0.90	0.67	0.90	0.67	HLA-C*14:02	0.67	0.28	0.67	0.29
HLA-B*15:02	0.77	0.53	0.77	0.53	HLA-C*15:02	0.79	0.52	0.80	0.55
HLA-B*15:03	0.90	0.75	0.90	0.75					

Table 2. Evaluation results of Kim's DCNN and DeepSeqPan

MHC	IEDB Ref	Measure Type	Count	AUC	
				Kim	DeepSeqPan (Binary)
HLA-A*01-01	1028282	t1/2	6	1.00	1.00
HLA-A*02-01	1026371	t1/2	34	0.70	0.76
HLA-A*02-01	1026840	Binary	341	0.85	0.85
HLA-A*02-01	1026840	IC_{50}	22	0.79	0.55
HLA-A*02-01	1026840	t1/2	22	0.69	0.58
HLA-A*02-01	1027079	Binary	15	0.80	0.80
HLA-A*02-01	1027471	Binary	43	0.79	0.88
HLA-A*02-01	1027588	Binary	18	0.70	0.82
HLA-A*02-01	1028285	t1/2	135	0.75	0.72
HLA-A*02-01	1028553	IC_{50}	22	0.85	0.95
HLA-A*02-01	1028554	IC_{50}	44	0.75	0.91
HLA-A*02-01	1028928	Binary	11	0.92	0.94
HLA-A*02-01	1029824	Binary	77	0.59	0.58
HLA-A*03-01	1028288	t1/2	221	0.84	0.85
HLA-A*03-01	1031253	IC_{50}	14	0.96	1.00
HLA-A*11-01	1026891	Binary	16	0.71	0.67
HLA-A*11-01	1028287	t1/2	219	0.79	0.76
HLA-A*24-02	1026840	Binary	346	0.84	0.83
HLA-A*24-02	1026840	IC_{50}	19	0.62	0.61
HLA-A*24-02	1026891	Binary	19	0.55	0.68
HLA-A*24-02	1028289	t1/2	423	0.73	0.74
HLA-A*30-01	1026840	Binary	347	0.87	0.85
HLA-A*30-02	1026840	Binary	360	0.73	0.72
HLA-A*30-02	1026840	IC_{50}	56	0.51	0.55
HLA-A*30-02	1026840	t1/2	56	0.48	0.49
HLA-A*31-01	315312	Binary	8	0.94	0.88
HLA-A*66-01	315312	Binary	16	0.39	0.14
HLA-A*68-01	1026840	Binary	436	0.86	0.84
HLA-A*68-01	1026840	IC_{50}	35	0.84	0.78
HLA-A*68-01	1026840	t1/2	35	0.42	0.33
HLA-B*07-02	1026371	t1/2	33	0.89	0.91
HLA-B*07-02	1026840	Binary	288	0.86	0.81
HLA-B*07-02	1028291	t1/2	136	0.82	0.82
HLA-B*07-02	1028553	IC_{50}	22	0.84	0.90
HLA-B*07-02	1028554	IC_{50}	52	0.80	0.80
HLA-B*07-02	1028928	Binary	11	1.00	1.00
HLA-B*07-02	1031253	IC_{50}	13	1.00	1.00
HLA-B*15-01	1028293	t1/2	570	0.74	0.62
HLA-B*15-02	1027131	Binary	14	1.00	1.00
HLA-B*27-05	1029125	Binary	21	0.97	0.95
HLA-B*27-05	1031253	IC_{50}	12	0.60	0.63
HLA-B*35-01	1028292	t1/2	363	0.81	0.74
HLA-B*35-01	1028554	IC_{50}	56	0.47	0.58
HLA-B*40-01	1026891	Binary	19	0.83	0.81
HLA-B*40-01	1026897	Binary	15	0.80	0.80
HLA-B*44-03	1028554	IC_{50}	46	0.54	0.80
HLA-B*57-01	1028554	IC_{50}	53	0.87	0.89
HLA-B*57-01	1029061	IC_{50}	17	0.90	0.95
HLA-B*58-01	1026840	Binary	433	0.87	0.85
HLA-B*58-01	1026840	IC_{50}	34	0.77	0.53
HLA-B*58-01	1026840	t1/2	34	0.44	0.58
HLA-B*58-01	1026891	Binary	20	0.66	0.67
HLA-B*58-01	1026897	Binary	22	0.81	0.90
HLA-B*27-05	1031959	Binary	13540	0.60	0.60