Convolutional architecture for prediction of peptide-MHC binding affinities

Rudolph Layko

National Research University – Higher School of Economics, Moscow, Russia

layko.ds@gmail.com

Vadim Nazarov

Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry RAS, Moscow, Russia

National Research University – Higher School of Economics, Moscow, Russia

vdm.nazarov@gmail.com

Mikhail Shugay

Pirogov Russian National Research Medical University, Moscow, Russia,

Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry RAS, Moscow, Russia,

Central European Institute of Technology, Masaryk University, Brno, Czech Republic,

Nizhny Novgorod State Medical Academy, Nizhny Novgorod, Russia

mikhail.shugay@gmail.com

Motivation. Immune system is one of the most essential part of any living organism consisting of great number of mechanisms and subsystems. One of the problem with high importance in immunoinformatics is prediction of binding affinities among major histocompatibility complex (MHC) molecules and their peptide ligands. Primary reason of such problem significance is its complexity owing to great polymorphism of MHC class I, which is in turn means major limitations in prediction techniques and wide range of applications in vaccine development because of potential possibility to analyze any MHC of interest.

Challenges. Through detailed overview of current state-of-the-art packages, including pan-allele NetMHCPan [1] and MHCflurry [2] it was investigated that they share same problems and constraints. The most challenging and significant one is low capability to generalize on unseen MHC due to their approach which is hard-wired on available MHC protein sequences during model training. Moreover, mentioned predictors have poor results on considerable number of MHC despite the fact of good performance on majority of alleles.

Proposed approach. In order to overcome aforementioned constraints we have developed and tested a wide spectrum of neural network architectures with inherent to all deepness comparing to architectures in works presented earlier. Following types of neural networks was trained and accurately validated: different architectures of recurrent neural networks (GRU [3], LSTM [4] and Bidirectional versions of this networks), Convolutional networks [5] and mixed architectures of this types. Basic justification for deepness was

complex structure of interaction between amino-acid sequences of peptide and MHC protein due to their high diversity. In order to compare performance of mentioned above predictors with our models we took peptide:MHC interaction data Immune epitope database (IEDB) [2] for homo sapiens species, constiting of approximately 100,000 MHC-peptide pairs for training and approximately 30,000 MHC-peptide pairs for testing. Our final model was inspired by ResNet architecture [6] and Wide ResNet [7] which achieved the state-of-the-art performance in image recognition tasks by introducing special technique of convolution through residual blocks. Input data for the proposed model is pairs of sequences of 20-dimensional embeddings for amino acids obtained from applying the word2vec [8] approach to human peptidome. Training of the model was proceeded with following approach: after getting embeddings for MHC and peptides, we put them in separate branches, each of them contains from one to three blocks with 1-dimensional convolutional filters of size 1. Sequence of blocks ends up with concatenation of output flattened vectors from conviolutional filters that follows up with full-connected layers for prediction of the binding affinity between input MHC and peptide.

Results. Our model show competitive results with the existent software packages such as NetMHCPan [1] and mhcflurry [2] (F1 score - 0.82, AUC - 0.9 for human peptides) and our model capable of prediction of binding affinities for unseen MHC sequences with low loss (F1 score - 0.72 and mean AUC - 0.84 for 5 most abundant MHC alleles), that we tested by leaving out a specific allele and training the model on the other alleles. Obtained results suggests usefulness of the proposed model and its applicability to the prediction of binding affinities for peptide:MHC complexes with no training data.

This work was supported by the Russian Science Foundation grant 17-15-01495.

1. Nielsen M, Lundegaard C, Blicher T, Lamberth K, Harndahl M, et al. (2007) NetMHCpan, a Method for Quantitative Predictions of Peptide Binding to Any HLA-A and -B Locus Protein of Known Sequence, *PLOS ONE 2(8): e796. https://doi.org/10.1371/journal.pone.0000796*

2. Rubinsteyn A., O'Donnell T., Damaraju N., Hammerbacher J. (2016); Predicting Peptide-MHC Binding Affinities With Imputed Training Data *bioRxiv doi: https://doi.org/10.1101/054775*

3. Chung J., Gulcehre C., Cho K., Bengio Y.(2014) Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling, arXiv:1412.3555

4. Zeyer A., Doetsch P., Voigtlaender P., Schlüter R., Ney H. (2016) A Comprehensive Study of Deep Bidirectional LSTM RNNs for Acoustic Modeling in Speech Recognition arXiv:1606.06871

5. Krizhevsky A., Sutskever I., Hinton G.(2012) ImageNet Classification with Deep Convolutional Neural Networks, *NIPS 2012* 4824:1097-1105.

6. He K., Zhang X., Ren S., Sun J. (2015) Deep Residual Learning for Image Recognition. arXiv:1512.03385

7. Zagoruyko S., Komodakis N. (2016) Wide Residual Networks arXiv:1605.07146

8. Mikolov T. et al. (2013) Distributed Representations of Words and Phrases and their Compositionality arXiv:1310.4546