Deep learning model for prediction of probability for thymic selection for T-cell receptor sequences

Sophya Tolstoukhova
National Research University – Higher School of Economics, Moscow, Russia,
sptolstoukhova@gmail.com

Evgenii Ofitserov
Tula State University, Tula, Russia
jineskimo@gmail.com

Mark Izraelson
Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry RAS, Moscow, Russia,
Pirogov Russian National Research Medical University, Moscow, Russia,
Nizhny Novgorod State Medical Academy, Nizhny Novgorod, Russia
mizraelson@gmail.com

Vadim Nazarov
Genomics of Adaptive Immunity Lab, IBCH RAS, Moscow, Russia,
National Research University – Higher School of Economics, Moscow, Russia
vdm.nazarov@gmail.com

Immune system is an essential part of human organism that supports its normal health condition and physiology by protecting it from different pathogens invading the body. Immune system functions in a complicated way with involvement of various cell interactions. T-lymphocytes (or T-cells) are one of the basic immunity weapons against virus infections. Each T-cell has an amino acid molecule on its surface called a T-cell receptor (TCR). A certain TCR is able to bind particular kind of pathogen peptide, this defines the uniqueness of a T-cell – two cells with similar receptors are clones. Due to the process of V(D)J-recombination – random reconstruction of genome parts, translated to TCR peptide later – the number of various TCRs achieves 10^{14} [1]. In the beginning of a life cycle T-lymphocytes undergo a selection during which possible autoimmune TCRs and TCRs with low efficiency filtered out. Thus selection impacts on the TCR diversity. With a development of sequencing technology it became possible to extract nucleotide sequences coding TCRs and use this data for computational experiments. In this work we build a deep learning model of the immune system selection and predict the probability of passing the selection by a certain TCR sequence. For this purpose we built different architectures of recurrent neural network (RNN) architectures such as Long

Short Term Memory (LSTM) [3] or Gated Recurrent Unit (GRU) [2] and tested them on 6 repertoires of 3 individuals with 2 replicas for every individual, using replicas as cross-validation datasets for corresponding individuals. To obtain sequences which are not passed the selection we create a dataset of 3,000,000 random sequences generated from the probabilistic model of V(D)J-recombination [1] and train the models on mixed dataset of selected and not selected sequences. Due to the random nature of generation process, some of the sequences in the generated datasets could possible pass the selection therefore we proposed an EM algorithm to relabel the data in the generated dataset during the training process. We also propose several approaches for data preprocessing and data generation in order to improve training results of the models. For each replica we tested our model on two datasets - for selected and filtered sequences. Out final model is capable to predict selected sequences with 0.97 accuracy and filtered sequences with 0.98 accuracy. Adding additional layers as well development of more advanced techniques such as variational autoencoders [4] could help lower the error. Our research findings show that bidirectional recurrent models works extremely better than their one-directional counterparts. Developed model could be used along with databases of antigen-specific TCR in order to predict whether antigen-specific TCR is presented in the dataset of interest.

This work was supported by the Russian Science Foundation grant 16-15-00149.

1. Murugan A., Mora T., Walczak A., Gallan C. (2012) Statistical inference of the generation probability of T-cell receptors from sequence repertoires, *PNAS* 109(40):16161-16166.

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

3. 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

4. Kingma D., Welling M. (2013) Auto-Encoding Variational Bayes arXiv:1312.6114