Applying deep-learning techniques to identification of pathogenic amino acid substitutions

Ivan Reveguk

St. Petersburg State University, 7/9 Universitetskaya emb., St. Petersburg, 199034, Russia,
edikedikedikedik@gmail.com

Ivan Sosin

St. Petersburg Academic University, 8/3 Khlopina Str, St. Petersburg, 194021, Russia,
iasawseen@gmail.com

Ilia Korvigo

Moscow Institute of Physics and Technology, Institutskiy per., 9, Dolgoprudnyy, Moscow Region, Russia
ilia.korvigo@gmail.com

Mikhail Skoblov

Moscow Institute of Physics and Technology, Institutskiy per., 9, Dolgoprudnyy, Moscow Region, Russia
mskoblov@gmail.com

Over the past decade deep learning techniques have achieved great success in such application as face recognition, text-mining, text-generation etc. Computational biology couldn't stay untouched. Due to dramatic progress in massively-parallel computation many problems in the field of computational biology are now gaining new perspectives. Among these is the functional analysis of protein molecules. It's well known that protein function can be dramatically affected by substitutions of single amino acids, and nowadays a lot of data regarding relationships between such substitutions and human diseases has been accumulated. Based on these data, we explore different deep-learning approaches that can be utilized in order to predict possible effects of single amino acid substitutions on protein function. We also focus on different ways one can represent protein structures, and how those can be used in two distinct neural network architectures in order to test different representations as well as in mentioned above classification problem.

It has already been shown in our work (Korvigo et al., 2017) that deep-learning methods can be successfully used to discriminate between neutral and pathogenic non-synonymous single nucleotide variations. We used massively-parallel training and a genetic algorithm to optimize

hyper-parameters. The resulting classifier outperforms all popular methods on several benchmarks and gains accuracy exceeding 90%. Hence we find it feasible to apply deep-learning to protein functional-space analysis.

According to published data recurrent neural networks (RNN) are a very promising technique for application in computational biology, e.g. predicting protein function (Søren Kaae Sønderby, 2014), gene regulatory networks (Khalid Raza, 2014) and protein secondary structures (Zhen Li, 2016). For our purposes we made use of layers with long short-term memory (LSTM) cells combined with convolutional layers. An LSTM is a recurrent network model that excels at remembering long- and short-time dependencies. We consider it a key advantage that can be used to predict phenotypical consequences of amino acid substitutions. Our goal was to infer possible damaging consequences of substitutions from protein sequence information, alone. We used the HUMSAVAR database as our source of reviewed categorized amino acid variations. For representation purposes we tried to encode the amino acids as one-hot encoded vectors, vectors of physico-chemical properties and vector-embeddings. The latter idea was inspired by recent research on continuous representation of molecules (Rafael G´omez-Bombarelli, 2016). We also tried to build a protein autoencoder similar to the molecular autoencoder (Rafael G ´omez-Bombarelli, 2016). Each amino acid was represented as one hot vector. Amino acid sequence for each protein was transformed into internal latent space. The goal was to recover amino initial sequence from latent space. We achieved accuracy up to 40 %.

In our second project we created a deep spatial convolutional neural network. CNNs can leverage spatial and temporal structure of the domain they model. It has been shown that CNNs can be applied, for example, to protein function classification problems (Zacharaki et al, n.d.), or to prediction of bioactivity of small molecules in ligand-protein interaction (Wallach, Dzamba, & Heifets, 2015). Since spatial structure is an intrinsic property of protein molecules, it is natural to expect that patterns of three-dimensional protein organization could be very prominent in predicting the functional impact of amino acid substitutions. We propose a voxel-based representation where each amino acid is represented by a non-empty voxel. In order to preserve important properties of a protein molecule which could serve as potential features in neural network training, we use two different ways of representing amino acids. The first one is based on the most important amino acid physical properties (Gu et al., n.d.; Li & Koehl, 2014). The

second approach is inspired by natural language processing techniques (Mikolov et al, n.d.) and utilizes vector embeddings. In our work, we also analyzed the sparsity of all protein models available at PDB and evaluated the information loss caused by down-scaling the representation. Based on these data we have selected an optimal representation size, which corresponds to maximum available memory capacity of GPU-units used during training process. We present our preliminary results on training CNN and CAE using model representations described above, discuss their advantages and limitations and emphasize differences in training results of chosen network architectures.

1. Korvigo, I., Afanasyev, A., Romashchenko, N., & Skoblov, M. (n.d.). Generalising Better: Applying Deep Learning to Integrate Deleteriousness Prediction Scores for Whole-Exome SNV Studies. https://doi.org/10.1101/126532

2. Søren Kaae Sønderby, Ole Winther (2014), Protein Secondary Structure Prediction with Long Short Term Memory Networks // arXiv.org:1412.7828v2

3. Khalid Raza, Mansaf Alam (2014), Recurrent Neural Network Based Hybrid Model of Gene Regulatory Network // arXiv.org:1408.5405

4. Zhen Li, Yizhou Yu (2016), Protein Secondary Structure Prediction Using Cascaded Convolutional and Recurrent Neural Networks // arXiv.org:1604.07176v1

5. Rafael G´omez-Bombarelli, David Duvenaud (2016), Automatic chemical design using a data-driven continuous representation of molecules // arXiv.org:1610.02415v2

6. Gu, S., Poch, O., Hamann, B., & Koehl, P. (2007). A Geometric Representation of Protein Sequences. In *2007 IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2007)* (pp. 135–142). IEEE. https://doi.org/10.1109/BIBM.2007.22

7. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality // arXiv.org:1310.4546

8. Wallach, I., Dzamba, M., & Heifets, A. (2015). AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery // arXiv.org:1510.02855

9. Zacharaki, E. I. (2017). Prediction of protein function using a deep convolutional neural network ensemble. https://doi.org/10.7287/peerj.preprints.2778v1