Drug-Target Interaction Prediction Toolkit

Prepared for SFL Scientific

The prediction of drug-target interactions is a crucial component of the drug development pipeline. As experimental approaches are costly, methods to predict these interactions *in silico* continue to garner significant attention in both the commercial and R&D sectors. In this vein, the objective of the SFL data scientist challenge is to design a machine learning approach to predict binding between protein/molecule pairs based on a supplied dataset with PubChem CIDs, UniProtKB IDs, and Kinase Inhibitors Bioactivity Data (KIBA) scores, along with an indication of whether or not the KIBA score was estimated or directly calculated.

The many challenges associated with creating such a tool can be condensed into three main areas: 1) acquisition of data, 2) featurization of inputs, and 3) feasible modeling strategies. Because the drug/target identifiers lack any descriptive quality, they must be converted into a meaningful representation of the drug/target while minding the workload to convert over 1 million records. Once converted, some relevant features must still be extracted from these representations in order to serve as input to a predictive model. Finally, the form and function of the model itself represent a trade off between performance and computational feasibility. A state-of-the art model, for instance, may theoretically provide excellent results but may be impossible to train with limited time and budget.

Towards overcoming the challenges outlined above, I have prepared a software repository containing a trained model which accepts the SMILES representation of a drug and the amino acid sequence of a target and directly estimates the KIBA score of the pair. My approach is modeled after the DeepDTA network proposed by Öztürk, H., et al., 2018. Rather than relying on handcrafted/extracted featurization of drugs and targets, this approach uses SMILES and protein sequences as input and learns to identify the most salient latent-features for predicting KBA. In short, the drug SMILES and protein sequences are converted to continuous, numerical representations, concatenated to form a combined representation for the drug-target pair, and used as input to a predictive model which attempts to predict KIBA score for the drug-target pair. This approach was selected as a means of solving issues 2) and 3), and to strike a balance between cutting edge (but computationally intractable) approaches, such as more modern transformer architectures, with classical approaches that require extensive handcrafting of drug/protein features. To solve issue 1) and therefore obtain the SMILES and protein sequences necessary to develop this model, I have prepared a set of utilities which leverage the public APIs for UniProt and PubChem. These utilities will collect the unique drug/target identifiers and submit a mapping job request to the PubChem and UniProt clusters, monitor the job status, and save the results to disk once ready. All code related to the development and training of the KIBA prediction model as well as a set of scripts for evaluating the performance of a test set are provided in the code repository for this project: https://github.com/nciovanac/DTA-Challenge

Significant opportunities yet remain for future work. While an internal validation set was utilized during training to help mitigate over-fitting, evaluation on the internal test set suggests that further model refinement is necessary. A true architecture and hyperparameter selection procedure through nested cross-validation could help to produce a more generalizable model. Access to additional computing resources, primarily CUDA capable GPUs, could facilitate faster model prototyping, the use of more modern deep-learning architectures, and could also provide the opportunity to investigate a wider scope of drug-target pairs. While significant development is still required, the foundational work conducted as a part of this challenge may hopefully aid in the development of a tool that can accelerate drug discovery workflows.