Team:Boun - DeepDTA: Deep Drug Target Binding Affinity Prediction

Hakime Öztürk[1], Elif Ozkirimli[2], and Arzucan Özgür[1] 1 Department of Computer Engineering, Bogazici University, Istanbul, Turkey 2 Department of Chemical Engineering, Bogazici University, Istanbul, Turkey

Abstract

For the IDG-DREAM Drug-Kinase Binding Prediction Challenge, we adopted a deep-learning based approach named DeepDTA [1] that was previously introduced by our team. DeepDTA aims to predict interaction strenghts of protein-compound pairs by utilizing Convolutional-Neural Network (CNN) blocks. These blocks learn high-level representations of proteins and compounds from their respective sequences. We employed DeepDTA with default parameter settings to predict binding affinities of Kinase-drug interactions. The source code for DeepDTA is available here: https://github.com/hkmztrk/DeepDTA

Introduction

The IDG-DREAM Drug-Kinase Binding Prediction Challenge aims to address the task of predicting affinities of the interactions between Kinases and drugs in terms of K_d (dissociation constant) values . The challenge consisted of three Rounds, Round1, Round1b, and Round2. We participated in Round1 and Round2. In both rounds, we used the DeepDTA model as the prediction system. In Round1, we obtained our best performance with the public dataset Davis [2] as training dataset, whereas in Round2, we obtained our best performance with the filtered subset of Drug Target Commons (DTC) dataset. Here we discuss the results of our best performance in Round2.

Methods

Training and test data

We used Drug-Target Commons (DTC) as our training data set. We filtered the original dataset based on the activity types that are related to K_d (e.g., pK_d, log K_d) and converted them to pK_d, values. After the filtration, we obtained total 50181 interactions between 1353 proteins and 11902 targets. We will refer to this dataset as (DTC_filtered) from now on. As for test dataset, the dataset provided for Round2 was used. Table 1 summarizes the training and test datasets.

Data	#interactions	#proteins	#drugs
Training	50181	1353	11902
Round2 Test	394	25	207

DeepDTA requires the sequence information of proteins and drugs in order to perform prediction. For both datasets, SMILES information for the drugs were available. Whereas for proteins, we utilized Python Bioservices [3] to collect protein sequences from the UniProt [4] database using respective UniProt IDs of the proteins.

Prediction Model

In this challenge, we adopted DeepDTA model [1] which is a Convolutional Neural Network (CNN) based architecture to predict binding affinities of drug-target interactions. The model consists of two respective CNN blocks, each learning high level representations from the sequences of proteins and drugs. The model combines these vectors into a single protein-drug representation. Finally, the concatenated representation is fed into a Fully-Connected Feed Forward Neural Network that consists of three layers with two dropout layers in between. Figure 1 below illustrates the architecture of the DeepDTA model.

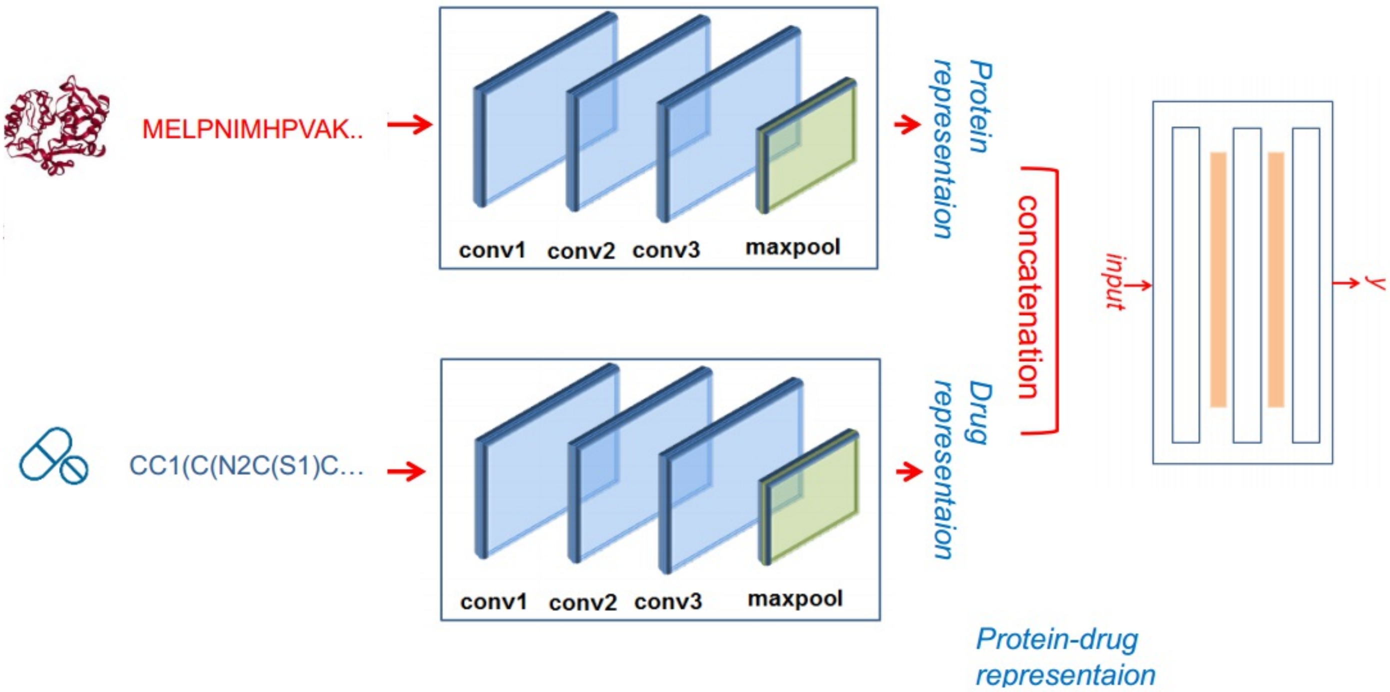

Figure 1: The summary of the DeepDTA architecture

We trained the model using the whole training data that we obtained after filtering (DTC_filtered) using the default parameter setting. Table 2 reports the default values of parameters that we utilized.

Parameter	value
Num. kernels	32, 64, 96
Protein kernel size	8
Drug kernel size	4
Batch size	256
Num. epoch	100
Dropout	0.1
FC	1024, 1024, 512

Keras [5] with Tensorflow [6] background is employed to build and train the DeepDTA model.

Conclusion

DeepDTA obtained values of 1.156 and 0.652 in terms of RMSE and AUC metrics in Round2 which scored average places in the Leaderboard. Since the model was trained with default values without any fine-tuning, the model achieves a promising results using only sequence information of proteins and compounds.

References

[1] Öztürk, Hakime, Arzucan Özgür, and Elif Ozkirimli. "DeepDTA: deep drug–target binding affinity prediction." Bioinformatics 34.17 (2018): i821-i829.

[2] Davis M.I.et al. . (2011) Comprehensive analysis of kinase inhibitor selectivity. Nat. Biotechnol. , 29, 1046–1051.

[3] Cokelaer, Thomas, et al. "BioServices: a common Python package to access biological Web Services programmatically." Bioinformatics 29.24 (2013): 3241-3242.

[4] Apweiler R.et al. . (2004) Uniprot: the universal protein knowledgebase. Nucleic Acids Res. , 32(Suppl. 1), D115–D119.

[5] Chollet F.et al. . (2015) Keras. https://github.com/fchollet/keras.

[6] Abadi M.et al. . (2016) Tensorflow: a system for large-scale machine learning. In: OSDI, Vol. 16, pp. 265–283.

Docker Instructions

Note: You have to place "input.csv" under the same directory as "Dockerfile"

```
docker build -t docker.synapse.org/syn18507647/deepdta:9686233 .

docker run -t -d -v  [your-dir]:/output
docker.synapse.org/syn18507647/deepdta:9686233

docker run -it --rm -d -v [your-dir]:/output -v  [your-dir]:/input
docker.synapse.org/syn18507647/deepdta:9686233
```