An ensemble method for drug-target interaction prediction

Yijingxiu Lu[1], Sangsoo Lim[2,4], Sungjoon Park[1], MinGyu Choi[4], Chang Yun Cho[3,4], Sun Kim[1,2,3,4]
[1]Department of Computer Science and Engineering, Seoul National University;
[2]Interdisciplinary Program in Bioinformatics, Seoul National University;
[3]BioInformatics Institute, Seoul National University; [4]AIGENDRUG, Ltd

Abstract

- Predicting the interactions between protein target and small molecular ligands is an essential part of drug discovery process.
- Recent years with the development of machine learning techniques, the identification of Drug-Target Interactions (DTI) is no longer dependent on costly, time-consuming traditional lab experiments.
- Protein kinases comprise the largest enzyme family, and their inhibitors are proving to be well tolerated for the treatment of cancer.
- In our work, we present a kinase-inhibitor interaction prediction ensemble model, EnsDTI, to aggregate model architectures with different encoding schemes, achieve better performance on diverse datasets.

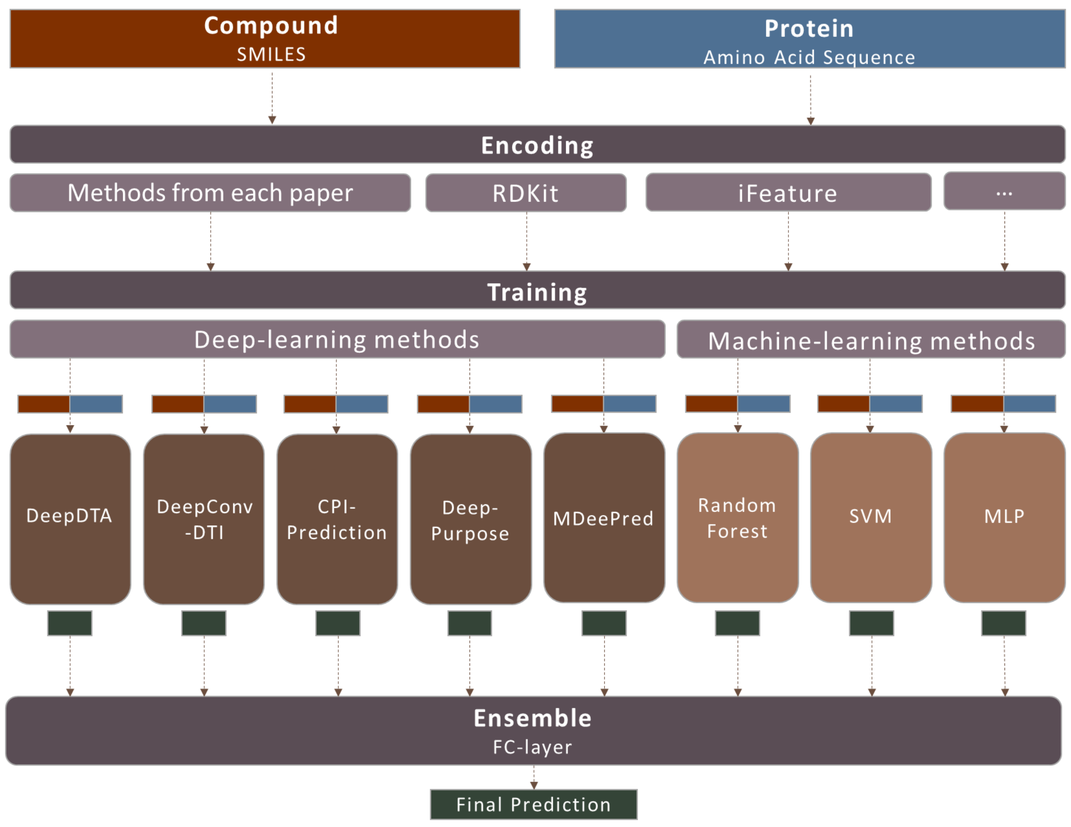

Figure 1. Overview of the ensemble model.

Motivation

- Although countless efforts have been made to predict drug target interactions, most of them only considered a small size of data space and used limited encoding techniques to represent protein target and small molecule ligand data.
- Biological and chemical materials can be represented with either 1D, 2D or 3D ways when considering different characteristics as the attributes of data to be learned with machine. A single model is hard to consider all perspectives that may affect interaction during data preprocessing.
- Currently presented methods have achieved not bad performance in the tasks of DTI predictions, but they can hardly be transferred into other types of ligand-target datasets as they are week in capturing distinct features from dissimilar data. Ensemble learning is considered as a feasible solution to overcome the shortage of weak generalization ability.

Method

- Using the framework in figure 1, we adopted five state-of-the-art methods (table 1) for DTI prediction to identify the inhibitors of protein kinases.
- An ensemble of above methods are presented to learn diverse binding mechanisms of general protein kinase inhibitors to improve generalization ability and accuracy, using different dimension of small molecular ligand and protein kinase target data.

Model	Task	Input		Ref
		Compound	Protein	
DeepDTA	Regression	SMILES	AA sequence	[1]
DeepConv-DTI	Classification	Morgan FP	AA sequence	[2]
CPI-Prediction	Classification	SMILES	AA sequence	[3]
DeepPurpose	Classification	SMILES	AA sequence	[4]
MDeePred	Regression	SMILES	AA sequence	[5]

Table 1. Deep-learning based classifiers

Interpretation Analysis

- In general, interpretation methods for machine learning architectures can be classified into two types – Gradient-based and Perturbation-based methods.
- Perturbation-based interpretation strategies that mask the input data to find the perturbed regions are usually applied when we only interested in the importance map of specific input queries instead of nodes or layers inside. These strategies are independent from the model to be explained and do not need to train additional modules.
- To provide explanation for EnsDTI model, we used perturbation-based interpretation method,, drawing the importance map of protein sequence to analysis the interaction contribution of each residue from the given protein (figure 2).

Figure 2. Interpretation analysis

Conclusion

- We compared our model with SOTA methods and traditional models on several DTI datasets with different evaluation metrics, and the results are shown in table 2. We can see from the results that our ensemble model performs better and more robust on various datasets, which indicate the effectiveness and generalization ability of this stacked ensemble framework for predicting drug-target interactions.
- Following the interpretation scheme introduced above, we draw the interaction contribution map of protein 4AN2 (figure 3). The reference map visualizes the actual binding sites of 4AN2. We can see from the result that the prediction of our ensemble method depends slightly upon the structure of binding site. Thus, the mechanism deserves further analysis for the model explanation.

		EnsDTI	Deep-learning methods				Machine-learning Methods		
			DeepConvDTI	DeepDTA	CPI Prediction	MDeePred	SVM	RF	MLP
Kiba (DeepDTA)	Acc	**0.906**	0.890	0.881	0.817	0.790	0.847	0.878	0.843
	Pre	**0.930**	0.913	0.918	0.815	0.790	0.886	0.886	0.854
	Spe	**0.728**	0.658	0.687	0.152	0.000	0.444	0.531	0.380
	Rec	0.953	0.951	0.932	0.992	**1.000**	0.954	0.970	0.965
ChEMBL (MDeePred)	Acc	**0.815**	0.767	0.740	0.643	0.670	0.704	0.787	0.738
	Pre	**0.781**	0.726	0.716	0.528	0.651	0.639	0.778	0.704
	Spe	0.867	0.840	0.856	0.490	**0.880**	0.791	**0.881**	0.843
	Rec	0.734	0.653	0.561	**0.878**	0.347	0.569	0.641	0.577
Human (Tsubaki)	Acc	**0.949**	0.906	0.893	0.907	0.763	0.898	0.896	0.897
	Pre	**0.944**	0.918	0.865	0.900	0.877	0.908	0.921	0.906
	Spe	**0.933**	0.905	0.825	0.825	0.892	0.893	0.911	0.889
	Rec	**0.963**	0.906	0.951	0.951	0.653	0.902	0.883	0.904

Table 2. Results

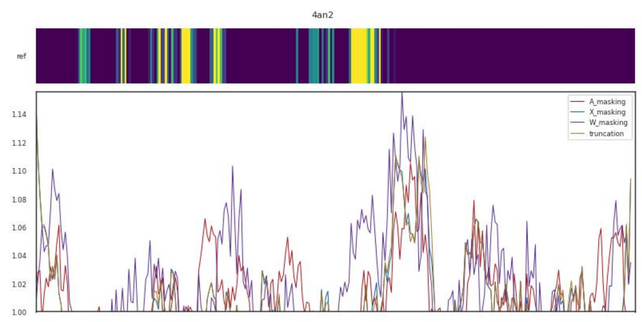

Figure 3. Visualization of interpretation analysis on protein 4AN2

References

[1] Öztürk H, Özgür A, Ozkirimli E. DeepDTA: deep drug–target binding affinity prediction[J]. Bioinformatics, 2018, 34(17): i821-i829.
[2] Lee I, Keum J, Nam H. DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences[J]. PLoS computational biology, 2019, 15(6): e1007129
[3] Tsubaki M, Tomii K, Sese J. Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences[J]. Bioinformatics, 2019, 35(2): 309-318.
[4] Rifaioglu A S, Cetin Atalay R, Cansen Kahraman D, et al. MDeePred: novel multi-channel protein featurization for deep learning-based binding affinity prediction in drug discovery[J]. Bioinformatics, 2020.
[5] Huang K, Fu T, Glass L M, et al. DeepPurpose: a deep learning library for drug–target interaction prediction[J]. Bioinformatics, 2020, 36(22-23): 5545-5547.