Deep and Shallow Chemogenomic Modelling for Compound-Target Binding Affinity Prediction Using Pairwise Input Neural Networks & Random Forests

Heval Ataş[1], Ahmet Rifaioglu[2], Tunca Doğan[1,3], Maria Jesus Martin[3], Rengul Atalay[1], Volkan Atalay[1,2]

[1] KanSiL, Department of Health Informatics Graduate School of Informatics, METU, Ankara, 06800 Turkey [2] Department of Computer Engineering, METU, Ankara, 06800 Turkey [3] European Molecular Biology Laboratory European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridge, CB10 1SD UK

Abstract

Machine learning techniques are frequently used in the field of drug discovery and repurposing for the prediction of interactions between drug candidate compounds and target proteins since the experimental approaches are not time- and cost-efficient to be applied to the massive compound-target interaction space. Recently, chemogenomic modelling approach became popular, where both compound and target protein features are used as inputs of the predictive models. Hence, they are able to incorporate targets with low number of (or no) training data and yield accurate predictions even for targets/compounds not involved in the training set at all. Chemogenomic approach is significant as it can be used to predict novel ligands for targets with limited training data and to identify the druggability potential of human proteins that were never targeted before. In this study, we developed chemogenomics-based computational methods, using random forest and deep neural network supervised learning techniques, to predict the binding affinities of a large set of kinases against several drug candidate compounds.

Introduction

The identification of binding affinity values between compounds and target proteins is critical for early stage drug discovery. Traditionally, binding affinity values are determined by high-throughput screening experiments, which are time-consuming and expensive, and thus, cannot be applied to the massive compound-target space. Therefore, computational methods have been developed to predict binding affinities, using machine learning (ML) techniques. Recently, chemogenomic modelling approaches became popular, where both compound and target protein features are used as the pairwise inputs of the predictive models. The output of these models are the binding affinity value predictions for the corresponding compound-target pair. The two main advantages of chemogenomic modeling are: (i) ability to incorporate targets with low number of (or none) training data points, (ii) potential to achieve elevated predictive performance due to more complex modeling. Here, we developed chemogenomics-based methods, on which, we tested different sets of compound and target protein features as input to observe the predictive performance.

Methods

In this study, we generated several predictive models (only 4 of them are shown here) for the binding affinity prediction of compound-kinase interactions using random forest (RF) and feed-forward pairwise input neural network (PINN) algorithms with different combinations of feature types and modeling approaches, as shown in Table 1. Unlike many conventional ML-based compound-target interaction prediction studies, where the prediction is usually based on binary classification as active or inactive, we generated regression models to predict the quantitative binding affinity values of compound-kinase interactions.

We represented compounds with ECFP4 fingerprints (diameter: 2), which is one of the most widely used feature type for compounds, and we represented proteins as pssm-based feature vectors (i.e., tri-gram-PSSM and k-separated-bigram-PSSM). POSSUM web-server was employed to generate the feature vectors. We obtained experimental bioactivity data points for kinases from the ChEMBL database for training, where we included all bioactivities containing a pChEMBL value (i.e., -log(IC50, EC50, Ki, Kd, Potency, …)). For our first model, we used the all kinase

interaction data points with 192,935 data points to train a single model. For our second model, we generated seven sub-models. Each sub-model was trained with the data points of a specific kinase sub-family such as: Agc, Camk, Cmgc, Ste, Tk, Tkl and others including 15,706, 13,251, 21,498, 4,165, 66,385, 10,470 and 29,982 data points, respectively. The aim here was to observe if family specific modeling increases the predictive performance. For Model 1 and 2, we used RF algorithm with tree number = 100 and max_features = 0.33. RF model takes a concatenated feature vector (compound + target) as input. The final part is a regressor, which predicts binding affinity for the input compound-target pair in terms of pChEMBL values.

For Model 3 and 4, we used pairwise input feed-forward neural networks (PINN) as a deep-chemogenomic neural network architecture. The network takes a pair of feature vectors for compounds and targets from disjoint input nodes simultaneously, following certain number of processing layers, latent representation of compound and target features are concatenated and further processed on more feed-forward layers. The output layer is a single node (a regressor), which predicts binding affinity for the input compound-target pair in terms of pChEMBL values. We used two hidden layers for both and compound target side of the network. After the concatenation of compound and target hidden layers, two additional hidden layers were used before output. We examined different hyper-parameters concerning the number of neurons at each layer (4096, 2048, 1536, 1024, 512, 128), learning rate (0.01, 0.001, 0.005, 0.0001) and dropout rate (0.6, 0.8) before finalizing the model.

We evaluated model performances by 5-fold CV and by external validation on the IDG DTI prediction challenge test dataset (i.e., the experimentally identified bioactivity measures between a selected set of kinases and compounds, these data point has not been recorded in any bioactivity databases such as ChEMBL or PubChem yet) using root mean squared error (RMSE), Pearson and Spearman correlations, and F1-score. For F1-score, the problem should be transformed to classification, for this, we determined an active/inactive predicted binding affinity threshold of pChEMBL = 7.

Results & Conclusion

Cross-validation results are given in Table 2. For Model 2, we reported weighted means of seven sub-models for each metric. Model performance comparisons are given below:

Model 1 vs. 2: the family-specific model outperformed the all-kinases model, which is an important outcome in terms of data selection and modeling approach. It is probable that the models trained with a more focused dataset (i.e., data points belong to the members of a kinase family) performs better, because different kinase families have different ligand interaction properties, and the model that contain all kinases at once cannot generalize the data at hand successfully.

Model 3 vs. 4: these models performed similarly, which indicates that the effect of the target feature type was minimal between k-sep-bigrams and trigrams features, which are similar in terms of the underlying representation logic, but very different in terms of dimensionality (k-sep-bigrams: 400, trigrams: 8,000). It is important to note that, we previously examined several more target feature types, and k-sep-bigrams and trigrams were selected based on those preliminary tests.

Model 1 vs. 3: RF models outperformed PINN models considering the cross validation results. As stated in the literature, the performance of deep neural network models are highly dependent on the selected hyper-parameters. Until this point, we could not scan a vast hyper-parameter space yet, we believe this is the main reason behind the observed performance difference

Information on IDG-DREAM Drug-Kinase Binding Prediction Challenge - Round2 Submission

RF Model (ObjectId: 9686327) For this submission we employed the methodology used in Model 1, as explained above; however, we reduced the training dataset to only the data points of the target kinases that are presented in the Round2 test dataset. The finalized training set was composed of 94,184 activity measurements between 61,603 compounds and 204 kinases.

PINN Model (ObjectId: 9686326) For this submission we employed the exact same methodology used in Model 3, as explained above.

Commands To Run the Docker Containers

RF Model (ObjectId: 9686327)

```
sudo docker run -it --rm -v $PWD:/input -v $PWD:/output
docker.synapse.org/syn18636383/crossbar_chemogenomic-modelling_rf:9686327
```

PINN Model (ObjectId: 9686326)

```
sudo docker run pinn-kinase-prediction-model
```

Authors Contribution Statement

HA, ASR, MJM, RA, VA and TD conceived the idea. HA and ASR performed the system construction. HA, ASR and TD performed all data analyses. MJM, RA, VA and TD supervised the overall study. All authors have revised and approved the this document.

Tables

Table 1: Characteristics of RF and PINN models.

Algorithm	Training set of each model	Protein Feature	Drug Feature
Model 1	RF	all kinases (1 model)	k-sep-bigrams
Model 2	RF	kinase families (7 sub-models)	k-sep-bigrams
Model 3	PINN	all kinases (1 model)	k-sep-bigrams
Model 4	PINN	all kinases (1 model)	trigram

Table 2. Predictive model performance results in 5-fold cross-validation.

Model name	RMSE	Pearson correlation	Spearman correlation	F1-score
Model 1 (RF)	0.64	0.87	0.87	0.85
Model 2 (RF)	0.63	0.87	0.87	0.86
Model 3 (PINN)	0.73	0.72	0.65	0.65
Model 4 (PINN)	0.73	0.72	0.64	0.65