Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Subject Section

# KinasepKipred: A Predictive Model for Estimating Ligand-Kinase Inhibitor Constant $(pK_i)$

Govinda KC [1], Md Mahmudulla Hassan [2], and Suman Sirimulla [1,2,3]*

[1]Computational Science Program, College of Science, The University of Texas at El Paso, El Paso, Texas, USA.
[2]Department of Computer Science, College of Engineering, The University of Texas at El Paso, El Paso, Texas, USA. and
[3]Department of Pharmaceutical Sciences, School of Pharmacy, The University of Texas at El Paso, Texas, USA.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

## Abstract

**Summary:** Kinases are one of the most important classes of drug targets for therapeutic use. Algorithms that can accurately predict the drug-kinase inhibitor constant $(pK_i)$ of kinases can considerably accelerate the drug discovery process. In this study, we have developed computational models, leveraging machine learning techniques, to predict ligand-kinase $(pK_i)$ values. Kinase-ligand inhibitor constant $(K_i)$ data was retrieved from Drug Target Commons (DTC) and Metz databases. Machine learning models were developed based on structural and physicochemical features of the protein and, topological pharmacophore atomic triplets fingerprints of the ligands. Three machine learning models [random forest (RF), extreme gradient boosting (XGBoost) and artificial neural network (ANN)] were tested for model development. RF model was finally selected based on the evaluation metrics on test datasets and used for web implementation.
**Availability:** GitHub: https://github.com/sirimullalab/KinasepKipred, Docker: sirimullalab/kinasepkipred
**Implementation:** https://drugdiscovery.utep.edu/pki
**Contact:** ssirimulla@utep.edu
**Supplementary information:** Supplementary data are available *Bioinformatics* online.

## 1 Introduction

Protein kinases play important roles in a wide range of diseases such as cardiovascular disorders, inflammatory diseases, gastrointestinal stromal tumors and cancer, and can serve as drug targets for therapeutic use (Fabbro (2015)). Kinase inhibitors that inhibit the activity of deregulated protein kinases form the largest class of new drugs approved for cancer treatment. The interaction is usually measured as binding affinity values in terms of dissociation constant $(K_d)$, inhibition constant $(K_i)$ and half-maximal inhibitory concentration $(IC_{50})$. We have developed a predictive model to estimate kinase-ligand $pK_i$ values . In this paper, we have used only the data corresponding to $(K_i)$ values for our dataset. Algorithms that predict drug-target associations (Yamanishi *et al.* (2008))(Li and Lai (2007)) and binding affinities were previously published by other researchers (Pahikkala *et al.* (2014))(He *et al.* (2017))(Öztürk *et al.* (2018))(Kundu *et al.* (2018)). However, to our best knowledge there are no algorithms that are specific to kinase $K_i$ predictions.

## 2 Methods

### 2.1 Datasets

Kinase-drug $K_i$ data was obtained from two publicly available databases (Tang *et al.* (2018); Metz *et al.* (2011)). Drug Target Interaction (DTI) dataset from the Drug Target Commons (DTC) by Tang et al. (Tang *et al.* (2018)) was used for the model development and initial testing of the models. Since we were focused only on kinases, only DTI pairs of 118 kinases and 5,983 compounds with $K_i$ values were used. Thus, our final data set contained 67,894 instances. All $K_i$ values were converted into molar units and then recorded as the negative decadic logarithm of them. Data obtained from Metz et al. (Metz *et al.* (2011)) was used as the external dataset for evaluating our models. We filtered out the Metz data to remove all the overlapped data with DTC. The external dataset contained 148 kinases with 240 compounds contributing to unique 17,258 DTI pairs (with $K_i$ values) which were not used in our training and test data.

### 2.2 Molecular features

**Protein features:** Protein features were generated using the Python-based tool named propy (Cao *et al.* (2013)). Features were sequence derived structural and physicochemical features from the amino acid sequence.
**Ligand features:** Toplogical pharmacophore atomic triplets fingerprints (TPATFP) features were generated for ligands using Perl scripts from the MayaChemTools (Sud (2016)). More detailed explanation of protein and ligand features are provided in the supporting information.

### 2.3 Model development

Models were developed mainly based on the grid search method with 5-fold cross validation. They were based on the Scikit-Learn machine learning library for Python (Pedregosa *et al.* (2011)). We used the 25% of data for the test set and 75% of the data for the training set. Three different machine learning models (random forest, extreme gradient boosting, and artificial neural network) were developed and we compared their performances using several evaluation metrics. More detailed explanation of these three models are available in the supporting information. We also compared grid search vs random search (Bergstra and Bengio (2012)) for our random forest model to estimate the efficiency of each method.

## 3 Results

The developed models were evaluated using several metrics such as root-mean-square-error (RMSE), Pearson correlation coefficient (R), Spearman correlation coefficient ($\rho$), concordance index (Con. Index), and area under the receiver operating characteristic curve (AUC-ROC). The table 1,2 and 3 show the scores obtained for the test and Metz data set (external test dataset) using three models. Among three different models, random forest was found performing best with R 0.887, $\rho$ 0.846, RMSE 0.475, Con. Index 0.854, and AUC 0.957 for the test data set and 0.769, 0.669, 0.503, 0.749, and 0.938 respective scores for the external data set. More details about the results and a comparative study can be found in the supporting information.

Table 1. Performance of RF model on the test and Metz datasets

| Metrics | Random Forest | |
|---|---|---|
| | Test dataset (95%) | Metz dataset (95%) |
| R | 0.887 (0.881, 0.893) | 0.769 (0.759, 0.779) |
| $\rho$ | 0.846 (0.840, 0.853) | 0.669 (0.659, 0.679) |
| RMSE | 0.475 (0.465, 0.486) | 0.504 (0.495, 0.513) |
| Con. Index | 0.854 (0.851, 0.858) | 0.749 (0.744, 0.755) |
| AUC | 0.957 (0.954, 0.960) | 0.938 (0.930, 0.946) |

Table 2. Performance of XGBoost on the test and Metz datasets

| Metrics | XGBoost | |
|---|---|---|
| | Test dataset (95%) | Metz dataset (95%) |
| R | 0.862 (0.855, 0.868) | 0.667 (0.656, 0.679) |
| $\rho$ | 0.795 (0.788, 0.802) | 0.574(0.563, 0.585) |
| RMSE | 0.522 (0.512, 0.533) | 0.582 (0.574, 0.591) |
| Con. Index | 0.805 (0.801, 0.810) | 0.703 (0.698, 0.709) |
| AUC | 0.946 (0.943, 0.950) | 0.922 (0.913, 0.931) |

Table 3. Performance of ANN on the test and Metz datasets

| Metrics | Artificial neural network | |
|---|---|---|
| | Test dataset (95%) | Metz dataset (95%) |
| R | 0.798 (0.794, 0.803) | 0.60 (0.592, 0.608) |
| $\rho$ | 0.716 (0.710, 0.722) | 0.498 (0.489, 0.507) |
| RMSE | 0.631 (0.623, 0.639) | 0.619 (0.612, 0.627) |
| Con. Index | 0.779 (0.774, 0.784) | 0.658 (0.651, 0.665) |
| AUC | 0.933 (0.926, 0.940) | 0.887 (0.880, 0.894) |

## 4 Web implementation and Code Availability

The model is available on a webportal at https://drugdiscovery.utep.edu/pki/. The web interface takes SMILES patterns and protein sequences as inputs, and provides the predicted results. Additionally, the model, data and results are available on github at github.com/sirimullalab/KinasepKipred. A docker image is also available via docker hub at sirimullalab/kinasepkipred

## Acknowledgements

We thank Drs. Mahesh Narayan, Amy Wagler, and Gabriel A. Frietze at UTEP for helpful discussions and Dr. Michael Scott Long for reading. We thank the High- Performance Computing Center at UTEP for assistance in using the Chanti cluster and the Texas Advanced Computing Center (TACC) at The UT Austin for providing HPC resources that have contributed to the research results reported within this paper.

## Funding

G.K was supported by the UTEP Computational Science Program. M.H and resources were supported by Dr. Sirimulla's startup fund from the UTEP School of Pharmacy..

## References

Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization. *JMLR*, page 305.

Cao, D.-S. *et al.* (2013). propy: a tool to generate various modes of Chou's PseAAC. *Bioinformatics*, **29**(7), 960–962.

Fabbro, D. (2015). 25 years of small molecular weight kinase inhibitors: Potentials and limitations. *Molecular Pharmacology*, **87**(5), 766–775.

He, T. *et al.* (2017). Simboost: a read-across approach for predicting drug-target binding affinities using gradient boosting machines. *J Cheminform*, **9**(1), 24–24. 29086119[pmid].

Kundu, I. *et al.* (2018). A machine learning approach towards the prediction of protein–ligand binding affinity based on fundamental molecular properties. *RSC Adv.*, **8**, 12127–12137.

Li, Q. and Lai, L. (2007). Prediction of potential drug targets based on simple sequence properties. *BMC Bioinformatics*, **8**, 353–353. 17883836[pmid].

Metz, J. T. *et al.* (2011). Navigating the kinome. *Nature Chemical Biology*, **7**, 200 EP –.

Pahikkala, T. *et al.* (2014). Toward more realistic drug–target interaction predictions. *Briefings in Bioinformatics*, **16**(2), 325–337.

Pedregosa, F. *et al.* (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

Sud, M. (2016). Mayachemtools: An open source package for computational drug discovery. *Journal of Chemical Information and Modeling*, **56**(12), 2292–2297.

Tang, J. *et al.* (2018). Drug target commons: A community effort to build a consensus knowledge base for drug-target interactions. *Cell Chemical Biology*, **25**(2), 224 – 229.e2.

Yamanishi, Y. *et al.* (2008). Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics (Oxford, England)*, **24**(13), i232–i240. 18586719[pmid].

Öztürk, H. *et al.* (2018). Deepdta: Deep drug-target binding affinity prediction. *Bioinformatics*, **34**.