

Software

Open Access



CrossMark

¹ Department of Biodiversity and Crop Improvement, International Center for Agriculture Research in the Dry Areas (ICARDA), Giza, Egypt.

² Department of Genome Mapping, Molecular Genetics and Genome Mapping Laboratory, Agricultural Genetic Engineering Research Institute (AGERI), Giza, Egypt.

³ Department of Bioinformatics Computer Networks, AGERI, Agricultural Research Center (ARC), Giza, Egypt.

Contacts of authors



* To whom correspondence should be addressed: Peter T. Habib

Received: September 24, 2020

Accepted: January 12, 2021

Published: January 20, 2021

Citation: Habib PT, Alsamman AM, Hassanein SE, Hamwih A. TarDict: A RandomForestClassifier based software predicts drug-target interaction using SMILES. 2021 Jan 20;1:bi202101




Copyright: © 2021 Habib *et al.*. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data availability statement: All relevant data are within the paper and supplementary materials.

Funding: The authors have no support or funding to report.

Competing interests: The authors declare that they have no competing interests.

TarDict: A RandomForestClassifier based software predicts drug-target interaction using SMILES

Peter T. Habib^{*1} , Alsamman M. Alsamman² , Sameh E. Hassanein³, Aladdin Hamwih¹ 

Abstract

The future of therapeutics depends on understanding the interaction between the chemical structure of the drug and the target protein that contributes to the etiology of the disease in order to improve drug discovery. Predicting the target of unknown drugs being investigated from already identified drug data is very important not only for understanding different processes of drug and molecular interactions but also for the development of new drugs. Using machine learning and published drug information we design an easy-to-use tool that predicts biological target proteins for medical drugs. TarDict is based on a chemical-simplified line-entry molecular input system called SMILES. It receives SMILES entries and returns a list of possible similar drugs as well as possible drug-targets. TarDict uses 20442 drug entries that have well-known biological targets to construct a prognostic computational model capable of predicting novel drug targets with an accuracy of 95%. We developed a machine learning approach to recommend target proteins to approved drug targets. We have shown that the proposed method is highly predictive on a testing dataset consisting of 4088 targets and 102 manually entered drugs. The proposed computational model is an efficient and cost-effective tool for drug target discovery and prioritization. Such novel tool could be used to enhance drug design, predict potential target and identify combination therapy crossroads.

Keywords: RandomForestClassifier, SMILES, drug-target interaction, Python, pathway.

Introduction

The identification of drug-target interactions (DTIs) leads to revolution drug discoveries. Drug developers search for drug compounds that interact with specific targets that has biological activities of interest. However, the identification of DTIs for enormous number of chemical compounds by experimenting usually takes 2-3 years, with high associated costs [1]. Thus, many computational methods and procedures developed to solve this problem. One of the most common computational methods, docking methods, which mimic the binding of a small molecule on a protein in 3D structure, were initially studied and used in all docking aspects. Docking methods try various scoring functions and molecule to decrease the free energy of binding. Docking methods have upgrade themselves, and currently, the Docking Approach introduce an alternate docking strategy termed DARC (Docking Approach using Ray-Casting), mapping the structure of a surface pocket “observed” from within the protein to the structure “observed” when viewing a potential ligand [2]. Moreover, many studies focused on similarity-based methods in which it was assumed that drugs bind to similar proteins and vice versa. Yamanashi *et al.*, used a kernel regression method to use the information on known drug interactions as the input to identify new DTIs, merging between a chemical genomic, and pharmacological approaches [3]. Those efforts successfully achieved the paradigm of 'one drug, one target' in the pharmaceutical field when it attracted attention to the role of the small number of main player genes interact with drugs [4]. This interaction shows how many drugs affect the body's proteins and explain how the development of the disease is often the result of a series of disruptions in our body's global pathway network environment [5].

Since it is time-consuming, expensive and requires considerable effort to be made to study different pathways and to determine whether a chemical and a pathway will interact with each other in a cellular network, it is reasonable to develop computational methods and machine learning algorithms to predict potential drug-target interactions in order to understand the drug mechanism of action [6].

One of the most widely used algorithms in bioinformatics is RandomForestClassifier, which has proved to be a model of choice for various machine learning projects worldwide. The Random Forest, as its name explains, consists of a number of decision-making trees that work together. Each tree in the random forest gives the forecast score and the most voted score chosen to be the prediction model [7,8].

In computational manner, there are many ways to deal with drug chemical structure, and Drug fingerprint is the most commonly used descriptor of drug substructure [9] where drug is transformed into a binary vector whose index value represents drug substructure existence. For proteins, descriptors are conventionally used as computational representations [10]. Unfortunately, feature-based models that use protein descriptors and drug fingerprints showed weak performance than conventional quantitative structure-activity relationship (QSAR) models [11]. Thus, We developed a computational prediction tool called "TarDict" that could be used to predict the biological target of any drug based on its SMILES string without needing for drug fingerprint. TarDict uses the RandomForestClassifier algorithm to construct a regression model to predict which target would be attacked by the drug's chemical compound. We used annotated structural drug data from DrugBank along with the target gene and pathway as a training data set to exceed our predictability accuracy.

Materials and Methods

Drug-target data

Drug-target data were obtained from the DrugBank database [12], which contains required information such as drug name, gene name, target pathway, and SMILES. We focused on 20442 SMILES of different drugs that have been tested and studied for liver carcinoma tissue and have well-known drug targets in the DrugBank database.

Machine Learning Analysis

Algorithm selection

To select the suitable algorithm to fit the data and be able to predict wisely, we build script that loop 23 algorithm under the same training and testing data. After testing and evaluation the mentioned algorithms, RandomForestClassifier was the best algorithm has accuracy and performance (**Code 1**).

Model building using RandomForestClassifier

A random forest is primarily a set of decision-trees, where each tree is a hierarchy of if/else questions which lead to decision-making. The only downside to decision-trees is their tendency

to overfit training data. each tree is different from the other in the random forest. The theory behind random forests is that any tree will predict fairly well but will probably overfit some of the data. Like other trees, all perform well and overfit in various ways, minimizing overfitness by means of an average of the results. This decline in overfitting, preserve the trees predictive ability.

Parameters tuning

A crucial parameter is tree number of forest that is determined by the parameter *n_estimators* in our script. We are constructing 30 trees. These trees are built separately and the algorithm allows specific random choices for each tree to ensure that the trees are distinct. However, the number of features chosen by each tree to create the if/else set is regulated by the *max_features* parameter of our example, which is 40. We called a bootstrap function for our data to make sure the forest is random. This is, we draw samples spontaneously with replacement from our data points frequently (the same data may be chosen many times). Random forest embedded in the machine learning library of Scikit-learn Python [13] (**Code 2**).

Figure 1 shows the random forest decision tree flowchart-like structure. We used the function CounterVectorizer, and it is programmatic aspects to Convert a collection of text to a matrix of token counts. We used CounterVectorizer to convert SMILES textual data to numeric to boost the learning ability of our machine learning model. We used the training file as control or/and standard to vectorize the SMILES input.

Using python programming language we packed the developed model into a standalone tool. TarDict receives SMILES entries and returns a list of possible similar drugs and possible targets. It connects the input drugs with biological pathways and eventually exports the possible pathway to the user. TarDict uses three steps to retrieve possible drug targets; (a) it receives drug SMILES information; (b) predicts the closest drug to this SMILES; and (c) RandomForestClassifier based model begins to identify the target that the predicted drug contributes to and finally exports the name of the pathway to the user.

Results and Discussion

Drug discovery and improvement are highly correlated with the information that could be obtained through known models of drug-target interactions. Where drug-target prediction tools could enhance therapeutics and increase the impact of medical research on human health. In this study, 20442 drug biological and chemical information collected from the DrugBank database was used to construct a machine learning model that could blindly predict the targeted pathway of novel drugs using chemical structure information.

Ensembles algorithms are the combination of several learning models, which enable more efficient models to be developed. Within machine learning science there are several models, but there are two ensemble models which are successful in a wide variety of datasets, all of which use decision trees as their build-

Code 1 [Python3]: Python script used for model validation.

```
1 algorithms = ['KNeighborsClassifier', 'GaussianNB',
2             'SGDClassifier', 'ExtraTreeClassifier',
3             'DecisionTreeClassifier', 'MLPClassifier',
4             'RidgeClassifierCV', 'RidgeClassifier',
5             'GaussianProcessClassifier', 'AdaBoostClassifier',
6             'GradientBoostingClassifier', 'BaggingClassifier',
7             'ExtraTreesClassifier', 'RandomForestClassifier',
8             'CalibratedClassifierCV',
9             'LinearDiscriminantAnalysis',
10            'LinearSVC', 'LogisticRegression',
11            'LogisticRegressionCV', 'NearestCentroid',
12            'Perceptron', 'QuadraticDiscriminantAnalysis',
13            'MultinomialNB']
14 for algo in algorithms:
15     model = eval('%s()' % algo)
16     model.fit(x_train, y_train)
17     y_pred = model.predict(x_test)
```

Code 2 [Python3]: Parameters used for randomforestclassifier library.

```
1 RandomForestClassifier(n_estimators=30,
↳ max_features=40, bootstrap=True)
```

Code 3 [Python3]: Importing countvectorizer module to be used for binarizing the SMILES strings.

```
1 from sklearn.feature_extraction.text import
↳ CountVectorizer
```

ing blocks: gradient-boosted and random forests that TarDict used.

To evaluate the predictive capability of TarDict, ten performance evaluation measures were applied. These predictive accuracy tests inform the different aspects of TarDict's performance (Table 1) using python script imports the module of each test. In addition, we validate the accuracy of the model prediction by testing the randomly selected SMILES data. Figure 2 shows the classification report assessing the predictive accuracy of TarDict for each drug class and the difference between the actual and the predictive target proteins.

Table 1. Accuracy test and performance evaluation measures results for TarDict constructed model using SMILES data set of drugs.

Parameter	Value
Accuracy	0.95
Balanced accuracy	0.94
Cohen kappa	0.95
Matthews corrcoeff	0.95
f1 score	0.95
hamming loss	0.05
Jaccard score	0.91
precision score	0.96
recall score	0.95
zero one loss	0.05
Time to build	0.3s

TarDict is an easy-to-use drug-target prediction where minimal information on the drug is needed to predict its potential biological target. This could facilitate therapeutic studies in which a large number of drugs are tested for their potential in targeting pathogenic genes. TarDict is an open-source software where researchers could use the same pipeline to create more effective prediction models for any drug class of interest. In addition, the process of predicting the closest drug will allow TarDict to be compatible with future drug findings and its potential use in drug classification by targetted biological pathways.

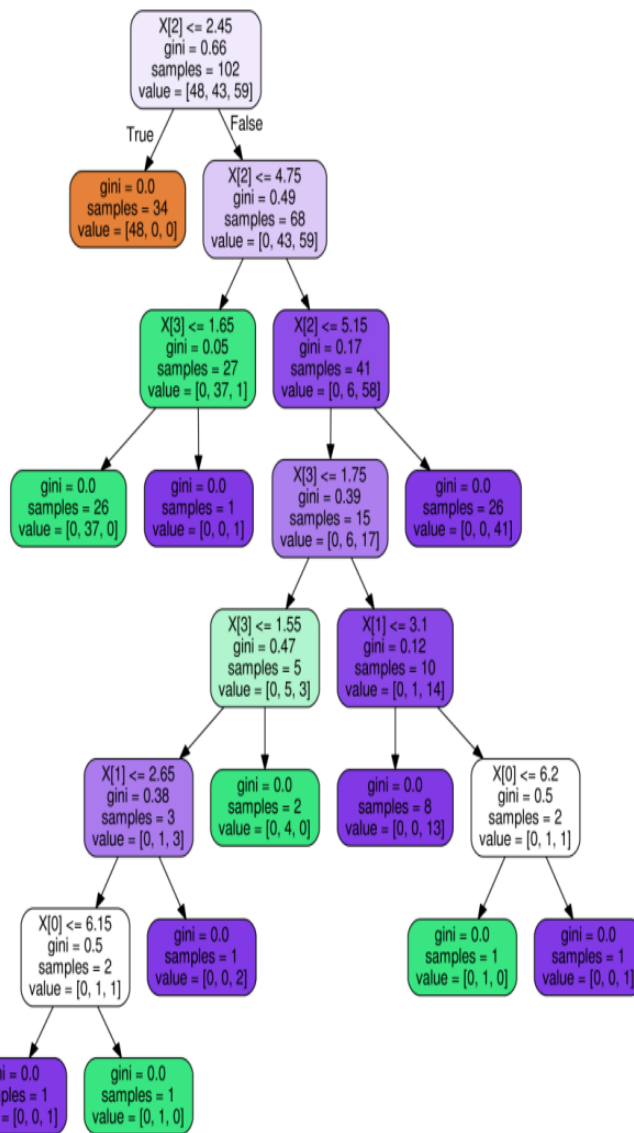


Figure 1. The constructed random forest decision tree flowchart-like structure for TarDict's pipeline.

Conclusions

In the drug target interaction area, the implementation of tools like TarDict is an outstanding innovation technology which evolves quickly with other modern fields of precision medicine, genomics and teleconsultation. This requires years and a million dollars to look for and grow therapeutic agents to cure a particular illness through clinical trials. Since approximately 17 databases and resources are available to detect drug-target interactions [13], TarDict's capacity to recognize the targets of an unknown drug is still dominant in this field. TarDict is an open source platform which could be modified to include additional features, modify the algorithm, and link its code with other tools.

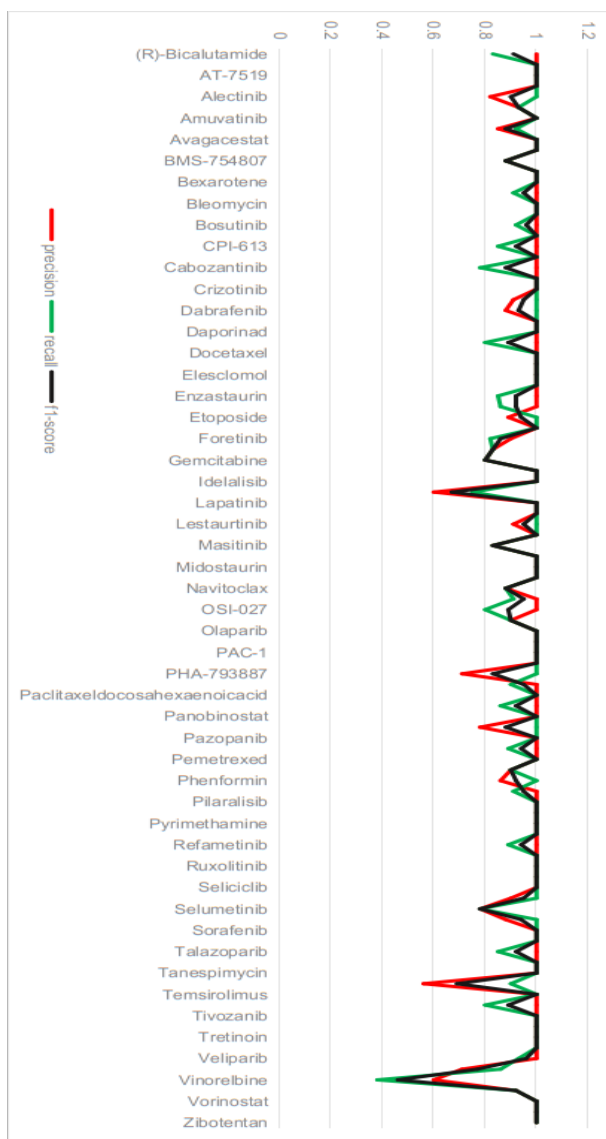


Figure 2. Classification report shows the prediction accuracy of TarDict using random SMILES test entry according to the precision the ratio and F1 score.

Availability

TarDict source code freely available on GitHub through <https://github.com/peterhabib/TarDict>.

References

1. Kapetanovic IM. Computer-aided drug discovery and development (CADD): in silico-chemico-biological approach. *Chemico-biological interactions*. 2008 Jan 30;171(2):165-76.
2. Gowthaman R, Miller SA, Rogers S, Khowsathit J, Lan L, Bai N, Johnson DK, Liu C, Xu L, Anbanandam A, Aubé J. DARC: mapping surface topography by ray-casting for effective virtual screening at protein interaction sites. *Journal of medicinal chemistry*. 2016 May 12;59(9):4152-70.
3. Yamanishi Y, Araki M, Gutteridge A, Honda W, Kanehisa M. Prediction of drugtarget interaction networks from the integra-

tion of chemical and genomic spaces. *Bioinformatics*. 2008 Jul 1;24(13):i232-40.

4. Ma H, Zhao H. iFad: an integrative factor analysis model for drug-pathway association inference. *Bioinformatics*. 2012 Jul 15;28(14):1911-1918.
5. Pujol A, Mosca R, Farrés J, Aloy P. Unveiling the role of network and systems biology in drug discovery. *Trends in pharmacological sciences*. 2010 Mar 1;31(3):115-23.
6. Yildirim MA, Goh KI, Cusick ME, Barabasi AL, Vidal Marc. Drug-target network. *Nat Biotechnol*. 2007;25(10):1119-26.
7. Masetic Z, Subasi A. Congestive heart failure detection using random forest classifier. *Computer methods and programs in biomedicine*. 2016 Jul 1;130:54-64.
8. Belgiu M, Drgu L. Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*. 2016 Apr 1;114:24-31.
9. Cereto-Massagué A, Ojeda MJ, Valls C, Mulero M, Garcia-Vallvé S, Pujadas G. Molecular fingerprint similarity search in virtual screening. *Methods*. 2015 Jan 1;71:58-63.
10. Dubchak I, Muchnik I, Holbrook SR, Kim SH. Prediction of protein folding class using global description of amino acid sequence. *Proceedings of the National Academy of Sciences*. 1995 Sep 12;92(19):8700-4.
11. Cheng F, Zhou Y, Li J, Li W, Liu G, Tang Y. Prediction of chemical-protein interactions: multitarget-QSAR versus computational chemogenomic methods. *Molecular BioSystems*. 2012;8(9):2373-84.
12. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic acids research*. 2018 Jan 4;46(D1):D1074-82.
13. Feurer M, Klein A, Eggenberger K, Springenberg J, Blum M, Hutter F. Efficient and robust automated machine learning. In: *Advances in neural information processing systems 2015* (pp. 2962-2970).