Applying Machine Learning Techniques for Classifying Cyclin-Dependent Kinase Inhibitors

Ibrahim Z. Abdelbaky[1]
Agricultural Research Center
Cairo,
Egypt

Ahmed F. Al-Sadek[2]
Agricultural Research Center, Cairo, Egypt, Computer Science Department, October University for Modern Sciences and Arts, MSA, October 6th, Egypt

Amr A. Badr[3]
Computer Science department, Faculty of Computers and Information, Cairo University
Cairo, Egypt

***Abstract*—The importance of protein kinases made them a target for many drug design studies. They play an essential role in cell cycle development and many other biological processes. Kinases are divided into different subfamilies according to the type and mode of their enzymatic activity. Computational studies targeting kinase inhibitors identification is widely considered for modelling kinase-inhibitor. This modelling is expected to help in solving the selectivity problem arising from the high similarity between kinases and their binding profiles. In this study, we explore the ability of two machine-learning techniques in classifying compounds as inhibitors or non-inhibitors for two members of the cyclin-dependent kinases as a subfamily of protein kinases. Random forest and genetic programming were used to classify CDK5 and CDK2 kinases inhibitors. This classification is based on calculated values of chemical descriptors. In addition, the response of the classifiers to adding prior information about compounds promiscuity was investigated. The results from each classifier for the datasets were analyzed by calculating different accuracy measures and metrics. Confusion matrices, accuracy, ROC curves, AUC values, F1 scores, and Matthews correlation, were obtained for the outputs. The analysis of these accuracy measures showed a better performance for the RF classifier in most of the cases. In addition, the results show that promiscuity information improves the classification accuracy, but its significant effect was notably clear with GP classifiers.**

***Keywords*—CDK inhibitors; random forest classification; genetic programming classification**

I. Introduction

Different important biological processes in the human body is related to the process of phosphorylation. In which, a phosphate group is added to proteins to activate their functionality. Protein kinases are enzymes that catalyze this process by adding the phosphate group to other proteins.

Due to the importance of the phosphorylation process in cellular processes and metabolism, protein kinases gained their importance and they are subject to many studies including drug design studies. Protein kinases are related to different diseases and cancer types when inappropriately regulated [1].

In humans, there are more than 500 kinases. They are divided into three types as they catalyze three types of phosphorylation [2].

Although kinases were targeted by several drug discovery studies that led to developing many inhibitors for them, only few of these inhibitors were approved. The reason for that is the undesired side effects caused by inhibitor reactivity against unintended targets. This is caused by the high degree of similarity between kinases, as there are few structural differences between them especially in their highly conserved binding domains. This similarity in binding domains led to the selectivity problem in many kinase inhibitors [2]. In most cases, this problem is caused by the high conservation is in the ATP binding site, which is the target for most of the inhibitors developed for kinases [3].

Among protein kinases, cyclin-dependent kinases (CDKs) are protein kinases, which have essential roles in cell divisions and transcription. CDKs are marked by being dependent on a protein subunit called cyclin to activate their enzymatic function. They belong to the serine/threonine kinase family [4].

Blocking the cell cycle by targeting kinases is proposed to kill cancer cells as in [5]. CDKs related to cell cycle are divided into three subfamilies, Cdk1, Cdk4, and Cdk5. The Cdk1 family consists of CDK1, CDK2, and CDK3 kinases. Although CDK1 is the most important kinase in this family because its major role for cell cycle, CDK2 is also essential as it participates in the cycle of cell division. In addition, CDK2 is investigated as being related to cancer and is targeted for cancer treatment as in [6].

CDK5 is an important enzyme that has different functions related to cell-cycle, gene expression, and others. CDK5 belongs to the cdk5 subfamily. In addition to its role in the cell-cycle progress, it is also known for controlling neuronal proteins [4]. CDK5 is also related to neurodegenerative diseases if was deregulated [7]. It is also linked to cancer and other diseases [8].

Computer-based approaches is being utilized in order to help profile the activity of different inhibitors against kinases and to explore and tackle the selectivity problem. Among these techniques is machine learning, which is widely utilized in biological and medical related problems. Different machine learning techniques were used in interaction modelling studies to predict protein-inhibitor interactions.

In [9] random forest was used to classify kinases variants in order to understand the relation with different diseases. The

classification was based on protein kinases sequence features. The resulting accuracy was 88%.

In the area of kinase inhibitors, machine learning was used in [10] to a study the kinase inhibitory data of [11] in order to model the prediction of interactions between kinases and their inhibitors. The study aimed for building a computational-experimental framework by using Kernel-based regression methods on molecular descriptors and fingerprints of kinase inhibitors. The predicted results were found correlated to kinase assays experimental results by 0.77.

In [12] Machine learning for predicting the binding of kinases to inhibitors by modelling different sets of features. Features used for kinases are based on sequences, in addition to phylogenetic features and amino acid positions in the active site. For inhibitors, 2D structural features and chemical features were used. Their experiments showed the importance of different sets of features based on the decision tree and SVM modelling results. The highest prediction accuracy achieved was 86.1%.

Another application of machine learning to predict active or inactive confirmations of kinases was done in [13]. The study proved that classification based on the activation segment orientation is performing better than other methods.

Genetic Programming (GP) [14] is a machine learning technique that simulates biological evolution and is used for modelling by regression or classification. It starts by a random population, then it continues to produce generations and individuals by performing evolutionary operations such as mutations, crossover, and selection, aiming to improve a fitness function. The individuals of GP is trees representing mathematical models to relate the modelled features to a target variable [15].

Random Forest (RF)[16], is a machine learning technique based on a large number of decision trees. A bootstrap sample is drawn and a set of variables are selected randomly to decide the split of each node. The tree grows and splits using the variables at each node until a specified criteria is achieved [17].

In this study, we use genetic programming and random forest classification techniques for classifying inhibitors and non-inhibitors for two of the cyclin-dependent kinases, CDK5 and CDK2. Both techniques were used for modelling chemical descriptors information. In addition to classification, we investigated the response of the classifiers to adding information about kinase binding promiscuity of compounds.

The outputs of the classifiers were analyzed using different accuracy measures and metrics. Because there is no standard single evaluator of classification accuracy, we calculated and obtained a group of measures for a wide evaluation of the results. These measures are confusion matrices, accuracy, ROC curves, AUC values, F1 score, and Matthews correlation coefficient.

Additionally, the analysis shows how could the classifiers reflect compound promiscuity information. Compound *promiscuity* against kinases is the ratio of the kinases that could be inhibited by that compound at a specific concentration [11].

This document is structured as follows: In section 2 we describe the dataset we used and illustrate data processing and workflow steps. In section 3, different results are presented and discussed. In section 3, we present our conclusion on the results and expectations for future improvements.

II. Data and Methods

We describe in this section the data sources, tools, and the methodology we used in order to achieve our objectives in building and evaluating the classifiers.

A. Data Sources

The dataset we used was extracted from the data of [11]. The original dataset contains the measured interaction values for more than 3000 compounds against 172 kinases. The values represent the pK_I values, which are the negative values of base-10 log of the K_I interaction value. We extracted the values for the first 1497 compounds against two protein kinases belonging to the cyclin-dependent kinases subfamily, CDK2, and CDK5.

The original dataset contains five cyclin-dependent kinases. We decided to study the data for CDK2, and CDK5 only as they have higher number of measured inhibitor activities with 868 and 1038 values respectively.

A threshold pK_I value of (value >5.9) was used for classifying compounds as inhibitors or not. This threshold was determined based on what the original study in [11] mentioned about compound activity against kinases.

We used the molecular descriptor values for the 1497 compounds. These values were obtained previously using e-dragon online tool [18]. The number of descriptors extracted for each compound is 1666 descriptor values.

Promiscuity value for each compound is provided in the original dataset in [11]. Promiscuity_1uM of a compound represents the portion of kinases tested with a potency of 1uM achieved by that compound.

B. Data Preparation

For each of the two proteins, two files were created with all the information needed for modelling. Each file contained the descriptor values for the compounds that interacted with one protein after removing columns that contained only zeros for all compounds. In addition, the interaction values were added as the last column as the target value. For each protein, another version of the data file was created including the value of *promiscuity_1uM* for each compound as an additional feature. So, each of the two proteins had two data files, and building a classifier was done twice for each protein. One time with descriptor values only, and another time with promiscuity and descriptor values.

The interaction values were classified based on the threshold mentioned in [11], considering the value of 5.9 pK_I as the inhibition threshold. The data file for each protein was modified replacing the interaction value with the class number. Class 1 represents that the corresponding compound is a potential inhibitor (pK_I > 5.9) while class 2 represents a non-inhibitor compound (pK_I <= 5.9). Table I shows the counts of

compounds as inhibitors or non-inhibitors according to the specified threshold in both protein datasets.

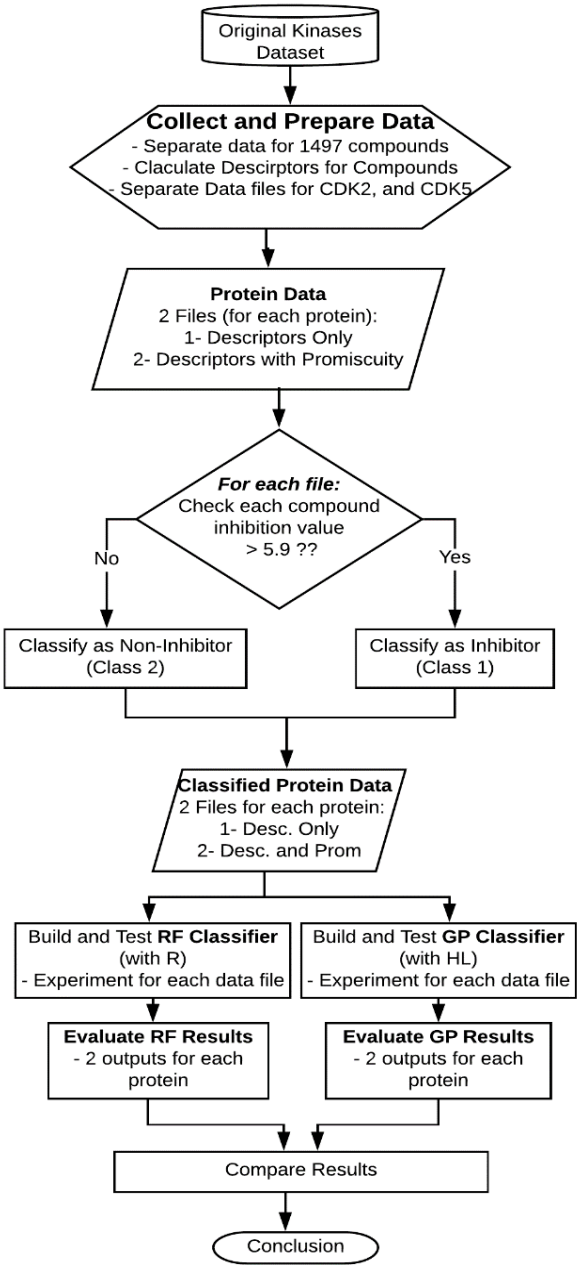


Fig. 1. Complete Workflow for Buidling and Eavaluating Classifiers.

Data separation, processing and cleaning were all done using python scripts that we wrote to read, manipulate and write csv files with the desired structure.

C. Methodology

In this study, we followed a multi-step methodology to explore the ability of two machine-learning techniques for predicting the inhibition activity of compounds against two cyclin dependent kinases. Random forest and genetic programming were applied on datasets of CDK5, and CDK2 kinases. The effect of adding promiscuity to the modelling features was also investigated, as it provides information about the ability of a compound to inhibit kinases.

First, we obtained, separated, and preprocessed the data sets for CDK5 and CDK2 kinases. After that, we obtained and prepared the required tools for testing random forest, and genetic programming classifications. Then, we performed different experiments, namely four for each protein, and collected the outputs. Finally, we evaluated the performance of the classifiers with different measures and compared the results. We concentrated more on the outputs of the RF classifier. The workflow of the complete steps for our work is shown in Fig. 1, which shows the steps followed to build both classifiers for each protein dataset.

In all experiments, descriptor values were considered as variables or features, and the class number was the target to be predicted. Each dataset was divided into a 70% training set, and a 30% test set.

D. Genetic Programming Classification

To perform GP classification we used a free desktop tool, HeuristicLab Optimizer 3.3.15 [19], in the mode (symbolic classification).

The input in each GP experiment was one of the files we created previously, in addition to setting GP parameters. We used different combination of parameters trying to achieve higher accuracy. The set of parameter values used with GP experiments are shown in table II.

TABLE I. Number of Compounds in ecah class

	Inhibitors	**Non-Inhibitors**
CDK5	234	804
CDK2	251	618

TABLE II. GP CLASSIFICATION PARAMETER VALUES

Population size	1000
No. of generations	1000 - 5000
Selection method	Tournament Selector
Crossover rate	0.90
Mutation rate	0.15
Objective	Min (MSE), Or Min (Penalty Score)
Model Depth	10
Model length	100

E. Random Forest Classification

Data files for each protein were loaded into R studio. Each dataset was divided by random sampling into a 70% training set, and a 30% testing set for validation.

The R package *(randomForest)* was used for the modelling. We set two basic RF parameters, the number of trees constructed *(ntree)*, and the number of randomly preselected features, or variables, in each tree *(mrty)*. We tried different values for these two parameters until we achieved a relatively low error value. Parameter values used for RF experiments are

shown for different datasets in table III. The parameter values used with the datasets including promiscuity are also shown in the table. The clear difference when using promiscuity in the dataset was achieving higher accuracies with lower number of trees.

Finally, results of different experiments were collected and different accuracy metrics were calculated for enhanced analysis.

TABLE III. RF CLASSIFICATION PARAMETER VALUES

	Promiscuity	**Trees** *(ntree)*	**Variables** *(mrty)*
CDK5	No	600	70
	Yes	450	110
CDK2	No	600	85
	Yes	450	65

III. RESULTS AND DISCUSSION

Both machine-learning classification techniques, genetic programming and random forest, were tested on two datasets for two cyclin dependent kinases and their inhibitors. Results varied between datasets and techniques. We mention the results in this section showing different accuracy measures we used, along with a discussion of the variations in these accuracies.

A. Accuracy

RF classifier could classify all test data. Table IV shows the overall accuracy of the RF classifier in terms of all correctly classified items in training and testing sets for both proteins. The accuracy is also shown when promiscuity was used in the data set.

B. Confusion Matrix

A confusion matrix is a table with a specific layout that is usually used to describe and visualize the performance of a classification algorithm. We show here the confusion matrices for each experiment.

The confusion matrices resulted from RF experiments are shown in tables V to VIII. Table V shows the matrix for the RF Result Accuracies for both proteins with and without promiscuity.

TABLE IV. RF RESULT ACCURACIES FOR BOTH PROTEINS WITH AND WITHOUT PROMISCUITY

	Promiscuity	**Accuracy**	
		Training	***Test***
CDK5	No	83.47 %	83.97 %
	Yes	84.44 %	85.26 %
CDK2	No	83.22 %	80.08 %
	Yes	85.53 %	88.51 %

In all confusion tables, the columns show the number of predicted items in each class (Inhibitor, Non-Inhibitor), and the rows display the actual items in each class. The results are shown for training and testing sets.

Table V shows results from CDK5 dataset without promiscuity, while table VI shows the results for CDK5 dataset including promiscuity information. Similarly, for CDK2 dataset, tables VII shows the results without promiscuity information, while table VIII displays the confusion matrix of CDK2 dataset that includes promiscuity values.

It should be noticed that the test sets were selected by random sampling, so, number of items in each class will not remain the same among different experiments.

The confusion matrices in all experiments show a high ability of the RF classifier to identify non-inhibitors. On the other hand, the ability to identify inhibitors is not in the same level. The reason for that could be the imbalance in data provided for the classifier, as most of the compounds in the data sets are already non-inhibitors as shown in table I.

TABLE V. RF CLASSIFIER CONFUSION MATRIX FOR CDK5 INHIBITORS (DESCRIPTORS ONLY)

		Predicted			
		Training		***Testing***	
		Inhibitor	*Non-Inhibitor*	*Inhibitor*	*Non-Inhibitor*
Actual	*Inhibitor*	52	109	28	45
	Non-Inhibitor	11	554	5	234

TABLE VI. RF CLASSIFIER CONFUSION MATRIX FOR CDK5 INHIBITORS (DESCRIPTORS AND PROMISCUITY)

		Predicted			
		Training		***Testing***	
		Inhibitor	*Non-Inhibitor*	*Inhibitor*	*Non-Inhibitor*
Actual	*Inhibitor*	57	103	30	44
	Non-Inhibitor	10	556	2	236

TABLE VII. RF CLASSIFIER CONFUSION MATRIX FOR CDK2 INHIBITORS (DESCRIPTORS ONLY)

		Predicted			
		Training		***Testing***	
		Inhibitor	*Non-Inhibitor*	*Inhibitor*	*Non-Inhibitor*
Actual	*Inhibitor*	82	87	36	46
	Non-Inhibitor	15	424	6	173

TABLE VIII. RF CLASSIFIER CONFUSION MATRIX FOR CDK2 INHIBITORS (DESCRIPTORS AND PROMISCUITY)

		Predicted			
		Training		***Testing***	
		Inhibitor	*Non-Inhibitor*	*Inhibitor*	*Non-Inhibitor*
Actual	*Inhibitor*	100	75	49	27
	Non-Inhibitor	13	420	3	182

C. ROC Curves

The Receiver operating Characteristics curve (ROC Curve) was plotted for all outputs to understand the ability of each classifier in discriminating between inhibitors and non-inhibitors. RF ROC curves were plotted using ROCR package in R [20], and are shown in Fig. 2, while ROC curves for GP experiments were obtained from HeuristicLab, and are shown in Fig. 3.

The curves show a fairly high ability for the RF classifier to label and determine the class for test data. For additional better understanding of the ROC curves, we show the values of AUC (Area Under the ROC Curve) for these ROC curves in table IX.

From the AUC values and the ROC curves, we can see that RF outperforms GP with both protein datasets, especially when promiscuity information exists. The AUC values also show a remarkable improvement when promiscuity information exists in the dataset for both proteins and with the two techniques. However, the improvement ratio in the case of promiscuity information is notably higher with GP classifier.

D. F1 Score

F1score is calculated based on the precision and recall measures. F1 score measures the accuracy of a classification model based on the number of positives identified correctly and the total number of positives. Tables X and XI show the F1 scores for RF and GP classifiers on CDK5 and CDK2 dataset respectively.

For CDK2, F1 scores are almost within a close range to each other except for GP classifier without promiscuity. Also in this measure, we can see that GP could better reflect the promiscuity information by increasing the F1 score value with a higher ratio than RF, although RF values were better beforehand. Additional note here is that CDK5 dataset without promiscuity could not result in high accuracy predictions of positives, even after many experiments.

E. Matthews Correlation Coefficient

Matthews correlation coefficient (MCC), is a quality measure used to evaluate binary classifications. So it is applicable in our case. It takes into consideration true positives and negatives and hence it is considered as a balanced measure. Tables XII and XIII show the MCC values for RF and GP classifiers on CDK5 and CDK2 datasets respectively.

TABLE IX. AREA UNNDER ROC CURVE FOR RF AND GP CLASSIFIERS

	AUC			
	CDK5		***CDK2***	
	No Prom.	*With Prom.*	*No Prom.*	*With Prom.*
RF	0.82	0.94	0.87	0.94
GP	0.56	0.85	0.62	0.68

TABLE X. F1 SCORES FOR CDK5 DATASET RESULTS

	F1 Score			
	Training		***Testing***	
	No Prom.	*With Prom.*	*No Prom.*	*With Prom.*
RF	0.46	0.50	0.53	0.57
GP	0.62	0.70	0.23	0.51

TABLE XI. F1 SCORES FOR CDK2 DATASET RESULTS

	F1 Score			
	Training		***Testing***	
	No Prom.	*With Prom.*	*No Prom.*	*With Prom.*
RF	0.62	0.69	0.58	0.77
GP	0.66	0.88	0.39	0.43

TABLE XII. MCC SCORES FOR CDK5 DATASET RESULTS

	MCC			
	Training		***Testing***	
	No Prom.	*With Prom.*	*No Prom.*	*With Prom.*
RF	0.45	0.49	0.50	0.56
GP	0.53	0.62	0.11	0.42

TABLE XIII. MCC SCORES FOR CDK2 DATASET RESULTS

	MCC			
	Training		***Testing***	
	No Prom.	*With Prom.*	*No Prom.*	*With Prom.*
RF	0.55	0.63	0.51	0.71
GP	0.54	0.83	0.18	0.33

The values of MCC measure in general ranges between -1 (No prediction), and 1 (Perfect prediction). In this case, the values for MCC in both datasets almost near to 0.5 or higher, except in the cases where GP classifier predicts the test sets for both proteins. As a general note, GP is performing better than RF in training data, but it cannot predict test sets accurately. On the other hand, RF is more accurate in predicting the test sets classes.

It is also clear from the tables that MCC value increases when promiscuity information is included in the datasets. Promiscuity information improved the accuracy of GP

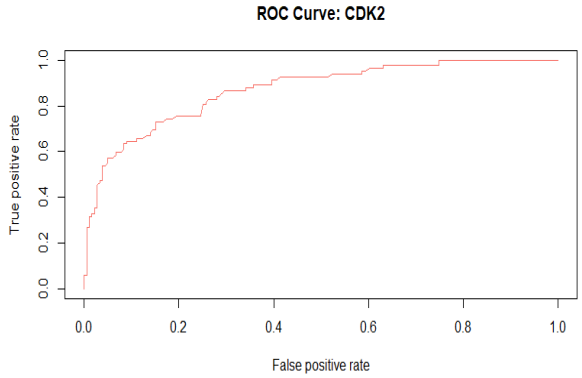
classifier more than its improvement for RF classifier on the training set level. This improvement is clearly noticeable in GP results for the test sets, although GP accuracy is still low on test sets compared to RF.

F. Important Vairables

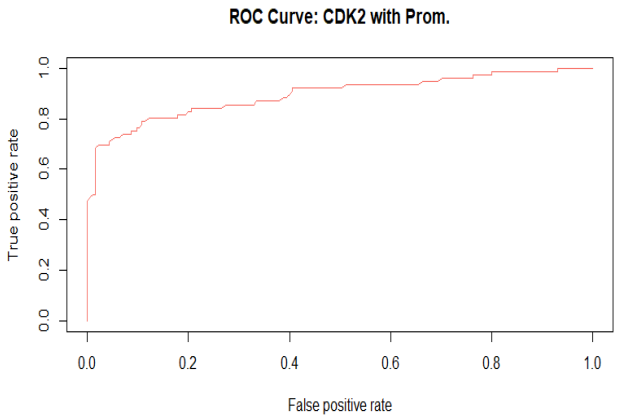
RF has the ability to rank different available considered while training. So, it can produce a list in each experiment with the most important variable affecting the prediction results. In table XIV we show a portion of the top important variables in the two experiments for each protein. We selected these important variables that had high rank in RF ranking for both mean decrease accuracy, and mean decrease Gini, and appeared with each protein in its corresponding two experiments. Variables names represent chemical descriptors produced by e-dragon.

TABLE XIV. IMPORTANT VARAIBLES AS SELECTED BY RF

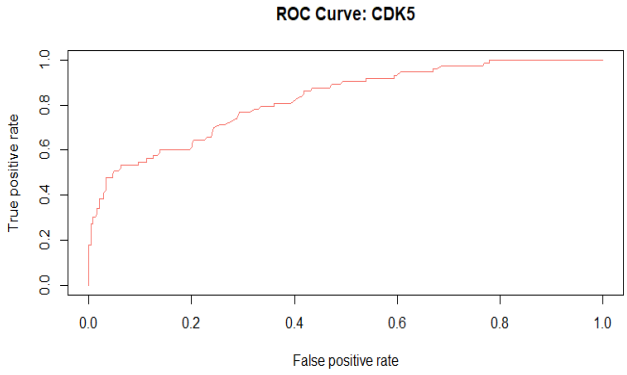
CDK5	CDK2
MATS7p	Mor32m
MAXDP	MATS1v
Cl-090	MATS1p
MATS7v	Mor18m



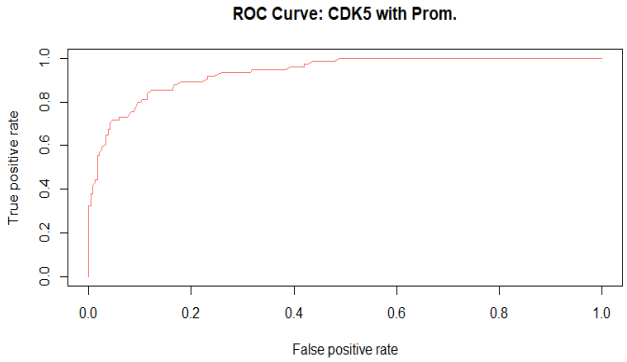
(a) CDK2 without Prmiscuity.



(b) CDK2 with Promiscuity.

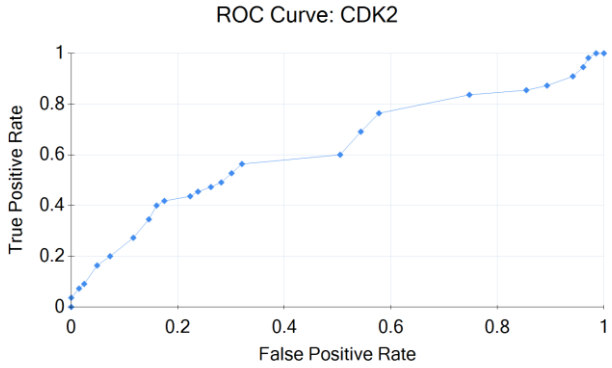


(c) CDK5 without Promiscuity.

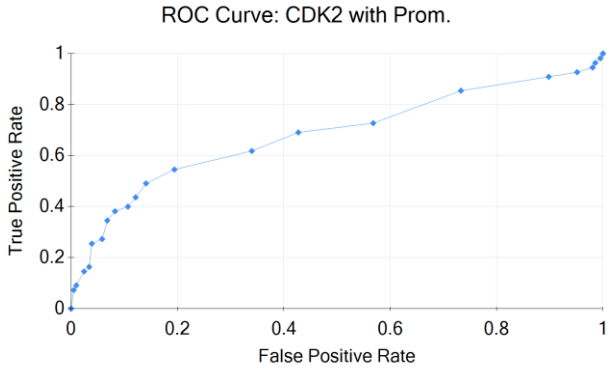


(d) CDK5 with Promiscuity.

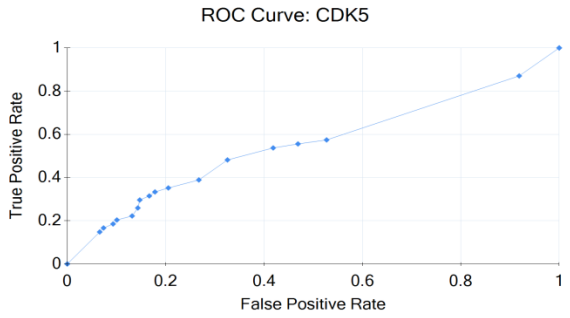
Fig. 2. RF Results ROC Curves for both proteins.



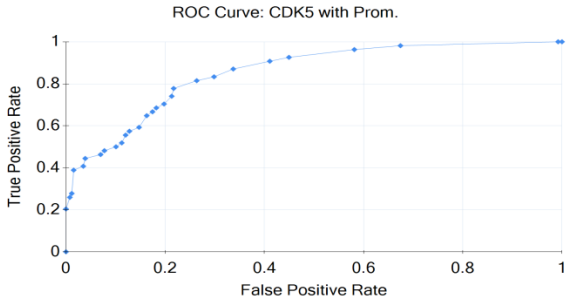
(a) CDK2 without Promiscuity.



(b) CDK2 with Promiscuity.



(c) CDK5 without Promiscuity.



(d) CDK5 with Promiscuity.

Fig. 3. GP Results ROC Curves for both Proteins.

IV. Conclusion

Machine learning techniques provides a useful means to model and understand kinase-inhibitor interaction data. Although the results were not usually of high accuracy for different accuracy measures, but still there are many measures showing promising values and representing good predictions. Machine learning classifiers produced good predictions for the class with more data in the dataset, non-inhibitor class. We suppose that this could be a result of imbalanced data distribution.

Another important conclusion is the ability of the classifiers to response effectively to one feature reflecting its importance. Kinase inhibitors are likely to bind to more than one kinase. The improvement in predictions when compound promiscuity is added to the features means that it was efficiently modeled. This suggests that adding more features such as protein binding site properties could highly improve the prediction accuracy.

Compared to previous work using different techniques mentioned in section 1, our results achieved promising values in terms of overall accuracy. The average overall accuracy from RF experiments was about 85%, which is comparable to the 88% in [9], and 86% in [12]. Most of previous work tried to predict kinase inhibitors for the whole family, while in our work here we concentrated on the CDK subfamily to be more specific and more responsive to any special binding features of CDKs. We expect that extending the data by adding more features and considering protein-related properties on different levels would improve the classification accuracy.

Finally, it is not necessarily that a good performing technique to be usually the most sensitive one for new features. Different approached should be tired with different datasets and features with a comprehensive and accurate evaluation of the results.

References

[1] S. Enna and D. Bylund, XPharm: The Comprehensive Pharmacology Reference. Amsterdam: Elsevier, 2008.

[2] H. Park, K. Kim, C. Kim, J. Shin and K. No, "Descriptor-Based Profile Analysis of Kinase Inhibitors to Predict Inhibitory Activity and to Grasp Kinase Selectivity", Bulletin of the Korean Chemical Society, vol. 34, no. 9, pp. 2680-2684, 2013.

[3] Y. Hu, N. Furtmann and J. Bajorath, "Current Compound Coverage of the Kinome", Journal of Medicinal Chemistry, vol. 58, no. 1, pp. 30-40, 2014.

[4] M. Malumbres, "Cyclin-dependent kinases", Genome Biology, vol. 15, no. 6, p. 122, 2014.

[5] W. Taylor and A. Grabovich, "Targeting the Cell Cycle to Kill Cancer Cells", Pharmacology, pp. 429-453, 2009.

[6] M. Sayour, I. Basheer, R. Salam, M. Yamany, A. Badr, A. Al-sadek and R. El-awady, "Potential mechanistic profiling of an otc analgesic as a cytotoxic agent in the treatment of hepatocellular carcinoma", ACTA Pharmaceutica Sciencia, vol. 56, no. 1, p. 95, 2018.

[7] Z. Cheung and N. Ip, "Cdk5: a multifaceted kinase in neurodegenerative diseases", Trends in Cell Biology, vol. 22, no. 3, pp. 169-175, 2012.

[8] A. Arif, "Extraneuronal activities and regulatory mechanisms of the atypical cyclin-dependent kinase Cdk5", Biochemical Pharmacology, vol. 84, no. 8, pp. 985-993, 2012.

[9] T. Pons, M. Vazquez, M. Matey-Hernandez, S. Brunak, A. Valencia and J. Izarzugaza, "KinMutRF: a random forest classifier of sequence variants in the human protein kinase superfamily", BMC Genomics, vol. 17, no. 2, 2016.

[10] A. Cichonska, B. Ravikumar, E. Parri, S. Timonen, T. Pahikkala, A. Airola, K. Wennerberg, J. Rousu and T. Aittokallio, "Computational-experimental approach to drug-target interaction mapping: A case study on kinase inhibitors", PLOS Computational Biology, vol. 13, no. 8, p. e1005678, 2017.

[11] Metz, E. Johnson, N. Soni, P. Merta, L. Kifle and P. Hajduk, "Navigating the kinome", Nature Chemical Biology, vol. 7, no. 4, pp. 200-202, 2011.

[12] F. Buchwald, L. Richter and S. Kramer, "Predicting a small molecule-kinase interaction map: A machine learning approach", Journal of Cheminformatics, vol. 3, no. 1, p. 22, 2011.

[13] D. McSkimming, K. Rasheed and N. Kannan, "Classifying kinase conformations using a machine learning approach", BMC Bioinformatics, vol. 18, no. 1, 2017.

[14] Koza, Genetic programming. Cambridge, Mass.: MIT Press, 1994.

[15] M. Tan, J. Tan, S. Chang, H. Yap, S. Abdul Kareem and R. Zain, "A genetic programming approach to oral cancer prognosis", PeerJ, vol. 4, p. e2482, 2016.

[16] L. Breiman, “Machine Learning,” Machine Learning, vol. 45, no. 3, pp. 261–277, 2001.

[17] A. Boulesteix, S. Janitza, J. Kruppa and I. König, "Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics", Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 2, no. 6, pp. 493-507, 2012..

[18] Tetko, J. Gasteiger, R. Todeschini, A. Mauri, D. Livingstone, P. Ertl, V. Palyulin, E. Radchenko, N. Zefirov, A. Makarenko, V. Tanchuk and V. Prokopenko, "Virtual Computational Chemistry Laboratory – Design and Description", Journal of Computer-Aided Molecular Design, vol. 19, no. 6, pp. 453-463, 2005.

[19] A. Elyasaf and M. Sipper, "Software review: the HeuristicLab framework", Genetic Programming and Evolvable Machines, vol. 15, no. 2, pp. 215-218, 2014.

[20] T. Sing, O. Sander, N. Beerenwinkel and T. Lengauer, "ROCR: visualizing classifier performance in R", Bioinformatics, vol. 21, no. 20, pp. 3940-3941, 2005.