JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

Article

pubs.acs.org/jcim

Binding Activity Prediction of Cyclin-Dependent Inhibitors

Indrajit Saha,[†,‡,§,○] Benedykt Rak,[†,○] Shib Sankar Bhowmick,[∥,⊥,○] Ujjwal Maulik,[∥]
Debotosh Bhattacharjee,[∥] Uwe Koch,[□] Michal Lazniewski,[†] and Dariusz Plewczynski*[,†,△,¶]

[†] Centre of New Technologies, University of Warsaw, 02-097 Warsaw, Poland

[‡] Institute of Informatics and Telematics, National Research Council, 56124 Pisa, Italy

[§] Institute of Computer Science, University of Wroclaw, 50-383 Wroclaw, Poland

[∥] Department of Computer Science and Engineering, Jadavpur University, Kolkata-700032, West Bengal, India

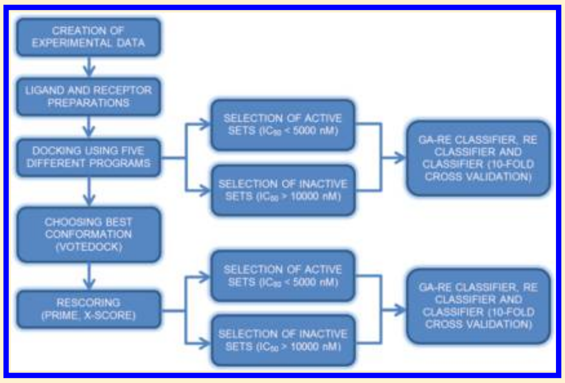
[⊥] Department of Informatics, University of Evora, Evora 7004-516, Portugal

[△] The Jackson Laboratory for Genomic Medicine, c/o University of Connecticut Health Center, Administrative Services Building−Call Box 901, 263 Farmington Avenue, Farmington, Connecticut 06030, United States

[¶] Yale University, New Haven, Connecticut 06520, United States

[□] Lead Discovery Center, Emil-Figge-Strasse 76a, 44227 Dortmund, Germany

Ⓢ *Supporting Information*


CREATION OF EXPERIMENTAL DATA
LIGAND AND RECEPTOR PREPARATIONS
DOCKING USING FIVE DIFFERENT PROGRAMS
CHOOSING BEST CONFORMATION (VOTEDOCK)
RESCORING (PRIME, X-SCORE)
SELECTION OF ACTIVE SETS (IC50 < 5000 nM)
SELECTION OF INACTIVE SETS (IC50 > 10000 nM)
SELECTION OF ACTIVE SETS (IC50 < 5000 nM)
SELECTION OF INACTIVE SETS (IC50 > 10000 nM)
GA-RE CLASSIFIER, RE CLASSIFIER AND CLASSIFIER (10-FOLD CROSS VALIDATION)
GA-RE CLASSIFIER, RE CLASSIFIER AND CLASSIFIER (10-FOLD CROSS VALIDATION)

ABSTRACT: The Cyclin-Dependent Kinases (CDKs) are the core components coordinating eukaryotic cell division cycle. Generally the crystal structure of CDKs provides information on possible molecular mechanisms of ligand binding. However, reliable and robust estimation of ligand binding activity has been a challenging task in drug design. In this regard, various machine learning techniques, such as Support Vector Machine, Naive Bayesian classifier, Decision Tree, and *K*-Nearest Neighbor classifier, have been used. The performance of these heterogeneous classification techniques depends on proper selection of features from the data set. This fact motivated us to propose an integrated classification technique using Genetic Algorithm (GA), Rotational Feature Selection (RFS) scheme, and Ensemble of Machine Learning methods, named as the Genetic Algorithm integrated Rotational Ensemble based classification technique, for the prediction of ligand binding activity of CDKs. This technique can automatically find the important features and the ensemble size. For this purpose, GA encodes the features and ensemble size in a chromosome as a binary string. Such encoded features are then used to create diverse sets of training points using RFS in order to train the machine learning method multiple times. The RFS scheme works on Principal Component Analysis (PCA) to preserve the variability information of the rotational nonoverlapping subsets of original data. Thereafter, the testing points are fed to the different instances of trained machine learning method in order to produce the ensemble result. Here accuracy is computed as a final result after 10-fold cross validation, which also used as an objective function for GA to maximize. The effectiveness of the proposed classification technique has been demonstrated quantitatively and visually in comparison with different machine learning methods for 16 ligand binding CDK docking and rescoring data sets. In addition, the best possible features have been reported for CDK docking and rescoring data sets separately. Finally, the Friedman test has been conducted to judge the statistical significance of the results produced by the proposed technique. The results indicate that the integrated classification technique has high relevance in predicting of protein−ligand binding activity.

1. INTRODUCTION

Cyclin-Dependent Kinases (CDKs) are members of the serine-threonine protein kinase family.[1] They are generally involved in controlling and sustaining the cell cycle. Moreover, CDKs are important for regulation of transcription as well as differentiation of nervous cells.[2] Current literature review shows that 11 CDKs have been identified and described. However, among them, only eight CDKs participate in cell cycle regulation.[3,44,45] Hence in this paper, we have confined ourselves to study those eight CDKs, named as CDK1, CDK2, CDK3, CDK4, CDK5, CDK6, CDK7, and CDK9. CDKs are regulated the cell cycle on the same way as on/off switch throughout current levels of cyclins, T-loop phosphorylation,[3] and in the presence of endogenous CDK inhibitors (CDKIs). Activation of a particular CDK allows the cell cycle to move into the next control point. On the contrary, deactivation results in halting the progression of cell cycle. The main cell cycle transitions require the activity of CDKs like the initiation of S-phase and starting of mitosis. Loss of CDK activity at the end of mitosis allows it to return to the interphase and is

Received: October 20, 2014
Published: June 16, 2015

DOI: 10.1021/ci500633c

Table 1. Cardinality of the Prepared Data Sets

activity	CDK1	CDK2	CDK3	CDK4	CDK5	CDK6	CDK7	CDK9
active	458	974	11	22	624	19	49	30
inactive	322	0	41	24	248	43	42	47

likely accompanied by dephosphorylation of the most target proteins that are modified when the cells entered to the mitosis. In this regard, it has been noticed that discovery of new CDKIs may lead to potential drugs for cell-cycle progression diseases, e.g., cancers.[4]

Molecular Docking (MD) is a method of molecular modeling that allows us to approximate the compound's position and conformation inside the receptor's active spot. It also tries to predict the binding affinity between two molecules.[5] It is commonly used in drug design because wet experiments are expensive and time-consuming. Roughly docking consists of three main stages. At first, it defines the binding site. There are few methods available in literature, e.g., homologous and discrete. Second phase involves optimization of the orientation as well as confirmation of the molecule inside the binding site. The last stage of docking uses scoring functions to calculate binding energies of complexes and create list of best hits. Here for our experiment, we have used several well-known docking programs, such as Surflex,[6] GOLD,[7] eHiTS,[8] C-Docker,[9] LibDock,[10] and Glide.[11]

The methods of finding best conformations contain old ones like rigid docking, which states ligands conformations do not change during docking process. Most of the current methods uses semiflexible docking techniques. The model of rigid docking is asymmetric, i.e., one molecule (normally the smaller ligand) is considered flexible, while the bigger one (receptor) is treated as rigid.[12] The four main semiflexible methods are Genetic tabu search algorithm,[13] Random docking,[14] Stochastic,[15] and Ligand fragmentation.[16] Among these methods, Random docking generates thousands of different ligands conformations and then dock them as rigid molecules (LibDock), whereas Stochastic approaches are used Monte Carlo based technique. As the stochastic docking program generates thousands of different ligand conformations, such generated ligands are then chosen by the Metropolis−Hastings algorithm[15] based on computed complex energy in order to accept or reject the conformers based on Boltzmann probability[15] (Glide). In ligand fragmentation, each ligand is divided into rigid fragments, where the biggest fragment is docked and the rest of the fragments are regrouped to dock as a single fragment. The rejoined parts are then minimized at the active site (Surflex, eHits).[17]

Scoring functions (SFs) are the mathematical methods, used to compute parameters between target molecule and ligand docked inside it. There are three types of SF such as force fields, empirical, and knowledge-based. In force fields SF, computation are based on interactions between atoms, i.e., van der Waals, electrostatic, and bonds parameters like length, strength, or angles.[18] Generally, the function properties came from quantum mechanics and experimental data. In empirical SFs, calculations are done based on approximation of binding affinity in energetic term as follows.

$$\Delta G = \sum_i (W_i \Delta G_i)$$

Where ΔG_i are different energy terms and W_i show multipliers of particular terms, which are determined using test sets of ligands with known binding affinity.[19] The knowledge-based SF uses the energy potentials obtained from structural data (delivered by experimental methods) and potentials are taken from the occurrence frequency of atomic pairs in 3D database, i.e., in Protein Data Bank.[20] Most of the current SFs are empirical-type.

The final stage of docking procedure using different scoring functions is called rescoring,[21] and it is more specified for calculating particular property, e.g., binding affinity. Stock scoring functions mainly focus on finding right conformation or rounded values (roughly approximated) of binding affinity. Moreover, two rescoring programs are using machine learning based regression model, named as X-Score[22] and MM GB-SA Prime[23] (Molecular Mechanics/Generalized Born Surface Area). X-score is pure empirical function based on 800 protein−ligand complexes, which is implemented by Wang et al.[22] as a standalone SF. On the other hand, Prime is a hybrid of force field and empirical scoring functions. It supports generalized Born model solvation (based on surface-area) with some empirical modifications in order to get appropriate solutions from the Poisson−Boltzmann equation.[24]

The values of these scoring functions can be used implicitly to capture binding activity of some unknown protein-peptide using machine learning based methods without modeling their complex binding structure. In this domain, for the prediction of CDK binders, a lot less research has been done to the best of our knowledge. For example, small set of ligands are experimentally validated only few specific kinase. In this regard, Kurapati et al. tried to use cross-docking to select CDK9 protein target using virtual screening.[48] On the other hand, Akrimah et al. merged MD with QSAR in order to discover in silico three new 4-(pyrazol-4-yl)-pyrimidine derivatives.[49] Despite the diversity of available various methods, there is no work fulfilling all of the requirements for predicting the binding activity between CDKs and ligands. This fact motivated us to develop a new prediction method with reliable improved classification technique.

To address the above issues, in this article, the Genetic Algorithm integrated Rotational Ensemble (GA-RE) based classification technique is proposed for predicting CDK binding activity. For this purpose, GA is used to encode the feature and ensemble size while the prediction accuracy is improved using a Rotational Feature Selection (RFS)[25,26] scheme with the integration of various machine learning techniques like Support Vector Machine (SVM),[27] Naive Bayes (NB),[28] Decision Tree (DT),[29] and *K*-Nearest Neighbor (*K*-NN),[30,31] separately. In RFS, GA selected features are subdivided into different nonoverlapping subsets to apply the Principle Component Analysis[32] technique on each subset in order to create a diverse set of features. Subsequently it is multiplied with the original training and testing data sets to generate new data sets to train and classify. This technique can be viewed as three levels of classification scheme, where level 1 deals with the traditional classifier while RFS and GA perform the tasks at levels 2 and 3, respectively. The classification performance of the GA-RE-Classifiers is validated through 10-fold cross validation for the different classifiers by measuring average accuracy, precision, sensitivity, specificity, F-measure, Matthews correlation coefficient (MCC), and area under the ROC curve (AUC) values

DOI: 10.1021/ci500633c

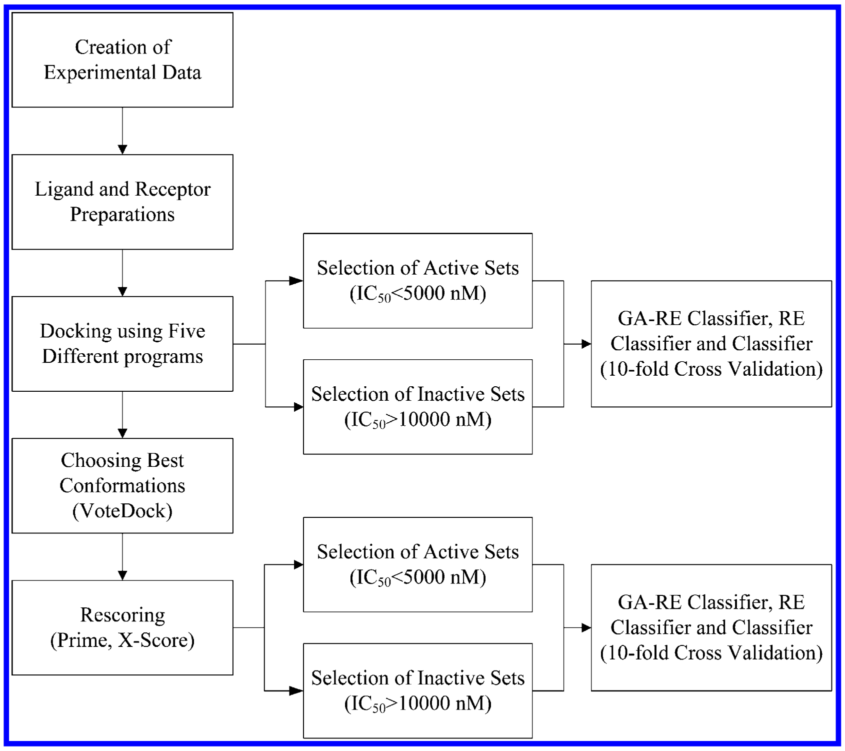

Creation of Experimental Data
Ligand and Receptor Preparations
Docking using Five Different programs
Choosing Best Conformations (VoteDock)
Rescoring (Prime, X-Score)
Selection of Active Sets (IC_{50}<5000 nM)
Selection of Inactive Sets (IC_{50}>10000 nM)
GA-RE Classifier, RE Classifier and Classifier (10-fold Cross Validation)
Selection of Active Sets (IC_{50}<5000 nM)
Selection of Inactive Sets (IC_{50}>10000 nM)
GA-RE Classifier, RE Classifier and Classifier (10-fold Cross Validation)

Figure 1. Schematic diagram of experimental procedure.

Table 2. Pearson Correlation Values of Experimentally Measured pIC50 and Docking Data

docking program	CDK1	CDK2	CDK3	CDK4	CDK5	CDK6	CDK7	CDK9
eHiTS	0.13	0.02	0.38	0.48	0.27	0.17	0.05	0.15
Glide	0.12	0.23	0.17	0.42	0.09	0.42	0	0.28
C-Docker	0.07		0.34	0.44		0.3		
LibDock	0.09	0.07	0.3	0.25	0.09	0.01	0.08	0.28
Surflex	0.14	0.14	0.28	0.14	0.15	0	0.16	0.05

Table 3. Pearson Correlation Values of Experimentally Measured pIC50 and Rescoring Data

rescoring function	CDK1	CDK2	CDK3	CDK4	CDK5	CDK6	CDK7	CDK9
Prime	0.01	0.02	0.39	0.07	0.09	0.23	0.09	0.15
X-Score	0.03	0.05	0.35	0.32	0.12	0.29	0.03	0.07

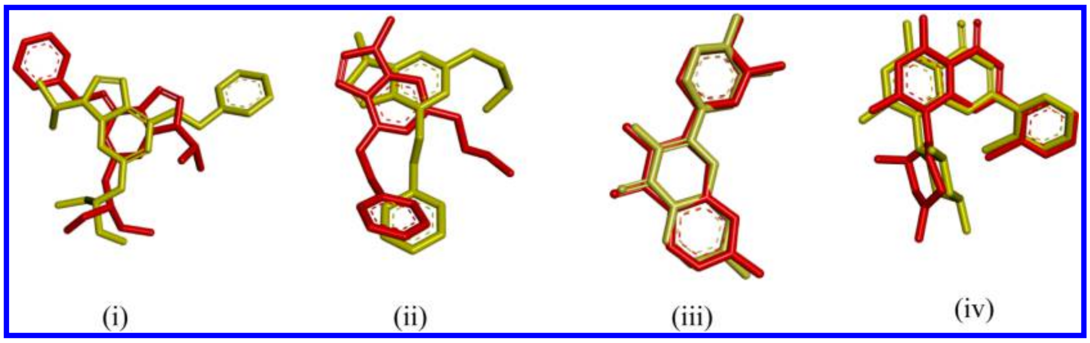


Figure 2. Alignment of exemplary VoteDock ligands compared to ligands from crystal structures (i) Roscovitine (CDK2), (ii) Olomoucine (CDK2), (iii) Fisetin (CDK6), and (iv) Flavopiridol (CDK9).

(see the Supporting Information for their definitions) for 8 CKDs in the form of 16 data sets of docking and rescoring, respectively. Finally, subset of potential features for each data set has been prepared, and the Friedman test[33,34] has also been conducted to judge the statistical significance of the results produced by the proposed technique.

2. MATERIALS AND METHODS

2.1. Preparation of Data Sets. During ligand preparation, a separate data set of ligands for each CDK has been prepared. Such ligands for CDK 2, 5, 7, and 9 are taken from University of Warsaw in SMILES format. Then the data sets are double checked in the PubChem[46] database, and duplicates are deleted.

Data sets of CDK 1, 4, 5, and 6 are collected manually from known public databases like PubChem[46] and BindingDB.[47] These data sets contain experimentally collected values of IC50, i.e., the half maximal inhibitory concentration. Generally, IC50 is the concentration of a compound required for 50% inhibition in vitro. All compounds are then translated from SMILES format to SDF and MOL2 format by Sybyl-X 2.0 unity application. Thereafter, ligands are refined by docking programs like Schrodingers LigPrep[11] and Accelryss Ligand Preparation.[9] Moreover, Surflex[6] and eHiTS[8] are used to prepare the ligands during docking procedure. Ligands for each kinase are divided into two sets, one contains only active compounds (IC50 < 5000 nM) and other contains only inactive ones (IC50 > 10000 nM). The cardinality of such data sets is mentioned in Table 1.

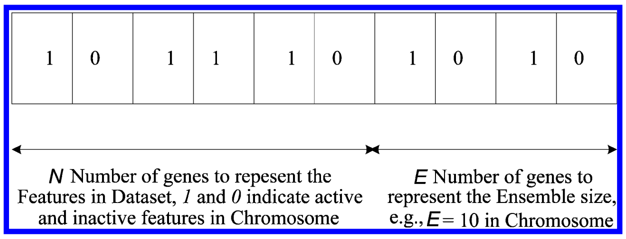

1 0 1 1 1 0 1 0 1 0
N Number of genes to repesent the Features in Dataset, *1* and *0* indicate active and inactive features in Chromosome
E Number of genes to represent the Ensemble size, e.g., *E* = 10 in Chromosome

Figure 3. Chromosome encoding scheme.

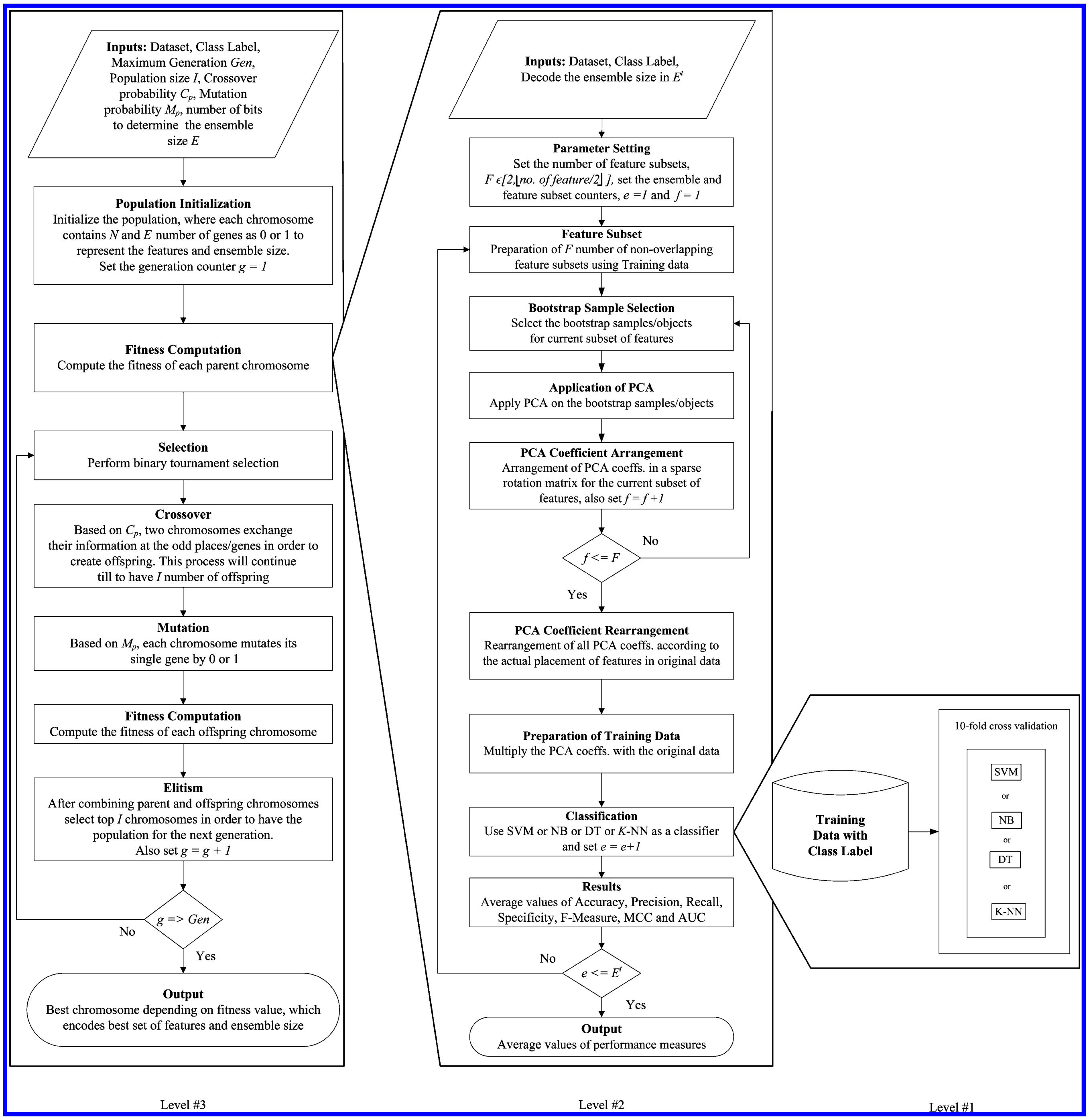

Inputs: Dataset, Class Label, Maximum Generation *Gen*, Population size *I*, Crossover probability C_p, Mutation probability M_p, number of bits to determine the ensemble size *E*
Population Initialization
Initialize the population, where each chromosome contains *N* and *E* number of genes as 0 or 1 to represent the features and ensemble size. Set the generation counter *g = 1*
Fitness Computation
Compute the fitness of each parent chromosome
Selection
Perform binary tournament selection
Crossover
Based on C_p, two chromosomes exchange their information at the odd places/genes in order to create offspring. This process will continue till to have *I* number of offspring
Mutation
Based on M_p, each chromosome mutates its single gene by 0 or 1
Fitness Computation
Compute the fitness of each offspring chromosome
Elitism
After combining parent and offspring chromosomes select top *I* chromosomes in order to have the population for the next generation. Also set *g = g + 1*
g => Gen
No
Yes
Output
Best chromosome depending on fitness value, which encodes best set of features and ensemble size
Inputs: Dataset, Class Label, Decode the ensemble size in *E′*
Parameter Setting
Set the number of feature subsets, *F* ϵ[2,⌊*no. of feature/2*⌋], set the ensemble and feature subset counters, *e =1* and *f = 1*
Feature Subset
Preparation of *F* number of non-overlapping feature subsets using Training data
Bootstrap Sample Selection
Select the bootstrap samples/objects for current subset of features
Application of PCA
Apply PCA on the bootstrap samples/objects
PCA Coefficient Arrangement
Arrangement of PCA coeffs. in a sparse rotation matrix for the current subset of features, also set *f = f +1*
f <= F
No
Yes
PCA Coefficient Rearrangement
Rearrangement of all PCA coeffs. according to the actual placement of features in original data
Preparation of Training Data
Multiply the PCA coeffs. with the original data
Classification
Use SVM or NB or DT or *K*-NN as a classifier and set *e = e+1*
Results
Average values of Accuracy, Precision, Recall, Specificity, F-Measure, MCC and AUC
No
e <= E′
Yes
Output
Average values of performance measures
Training Data with Class Label
10-fold cross validation
SVM
or
NB
or
DT
or
K-NN
Level #3
Level #2
Level #1

Figure 4. Block diagram of the proposed classification method.

Table 4. Average Accuracy Values of Different Classifiers for CDK Docking Data Sets

	Results of Docking Data based on the Average of 20 Runs											
	GA-RE Integrated Classifier (Level 3)				RE Integrated Classifier (Level 2)				Classifier (Level 1)			
Data Set	SVM	NB	DT	*K*-NN	SVM	NB	DT	*K*-NN	SVM	NB	DT	*K*-NN
CDK1	**92.47**	89.11	92.19	88.78	**90.19**	87.51	85.76	84.74	**88.14**	88.00	81.14	85.64
CDK2	93.10	96.89	**98.10**	97.99	89.73	**91.14**	90.73	89.13	85.71	84.25	**89.96**	82.24
CDK3	**87.00**	86.95	84.84	86.25	**84.88**	82.64	82.22	78.55	**81.98**	81.08	80.89	75.48
CDK4	87.54	83.78	**92.09**	88.88	**82.11**	78.21	76.09	72.13	69.58	73.72	**75.08**	70.60
CDK5	74.53	72.78	77.54	**80.08**	70.09	72.34	70.59	**72.44**	**72.44**	70.74	71.04	72.26
CDK6	**81.11**	75.09	80.88	81.08	**72.28**	71.71	70.18	70.18	70.73	**73.18**	70.98	72.48
CDK7	**82.56**	77.52	81.28	71.68	75.74	70.25	**77.86**	70.45	**72.84**	72.70	70.26	70.07
CDK9	**88.71**	75.25	81.17	81.83	**83.47**	74.56	79.89	80.14	**83.23**	71.25	74.78	75.56

Table 5. Average Accuracy Values of Different Classifiers for CDK Rescoring Data Sets

	Results of Rescoring Data based on the Average of 20 Runs											
	GA-RE Integrated Classifier (Level 3)				RE Integrated Classifier (Level 2)				Individual Classifier (Level 1)			
Data Set	SVM	NB	DT	*K*-NN	SVM	NB	DT	*K*-NN	SVM	NB	DT	*K*-NN
CDK1	**90.25**	90.10	90.15	89.55	88.07	**89.13**	85.03	86.20	68.15	78.02	**84.11**	77.53
CDK2	94.75	97.01	**97.59**	96.02	86.01	92.85	**95.33**	94.87	85.06	**92.57**	93.56	92.57
CDK3	86.33	89.60	86.33	**92.89**	81.01	86.14	85.23	**87.19**	80.03	**83.52**	82.53	79.71
CDK4	**94.85**	84.81	83.17	84.61	**81.63**	78.52	73.51	81.13	**76.14**	**76.14**	71.53	73.57
CDK5	82.65	75.35	**88.45**	84.71	**75.21**	74.72	71.45	74.53	**72.85**	72.05	70.03	72.45
CDK6	83.12	81.15	88.85	**92.53**	82.25	82.25	**86.31**	80.49	76.71	80.07	**81.55**	75.51
CDK7	**95.57**	92.11	92.11	92.11	86.22	84.79	**90.51**	87.61	71.46	80.01	**90.03**	83.62
CDK9	**93.25**	82.15	92.51	85.15	79.91	77.62	**89.77**	78.51	70.77	75.45	**80.52**	76.06

The crystal structures of each receptor excluding CDK1 have been downloaded from Protein Data Bank and refined later in each docking program for its compatibility, e.g., adjusting atom types, fixing side chains, etc. Note that CDK1 is created using Homology modeling from CDK2 template, which shares approximately 86% amino-acid identity[43] by I-TASSER software.[35]

The docking process is divided into two stages. First for CDK 2, 5, 7, and 9, both active and inactive, have been docked by four docking programs, Surflex Geom-X, Glide with extra precision mode, LibDock and eHits. In this stage, C-Docker has not been used. Second stage includes docking of CDK 1, 3, 4, and 6 for both active and inactive data sets. The fifth docking program, C-Docker, is added. Ligands for C-Docker have also been prepared by Accelrys's Ligand preparation tool.

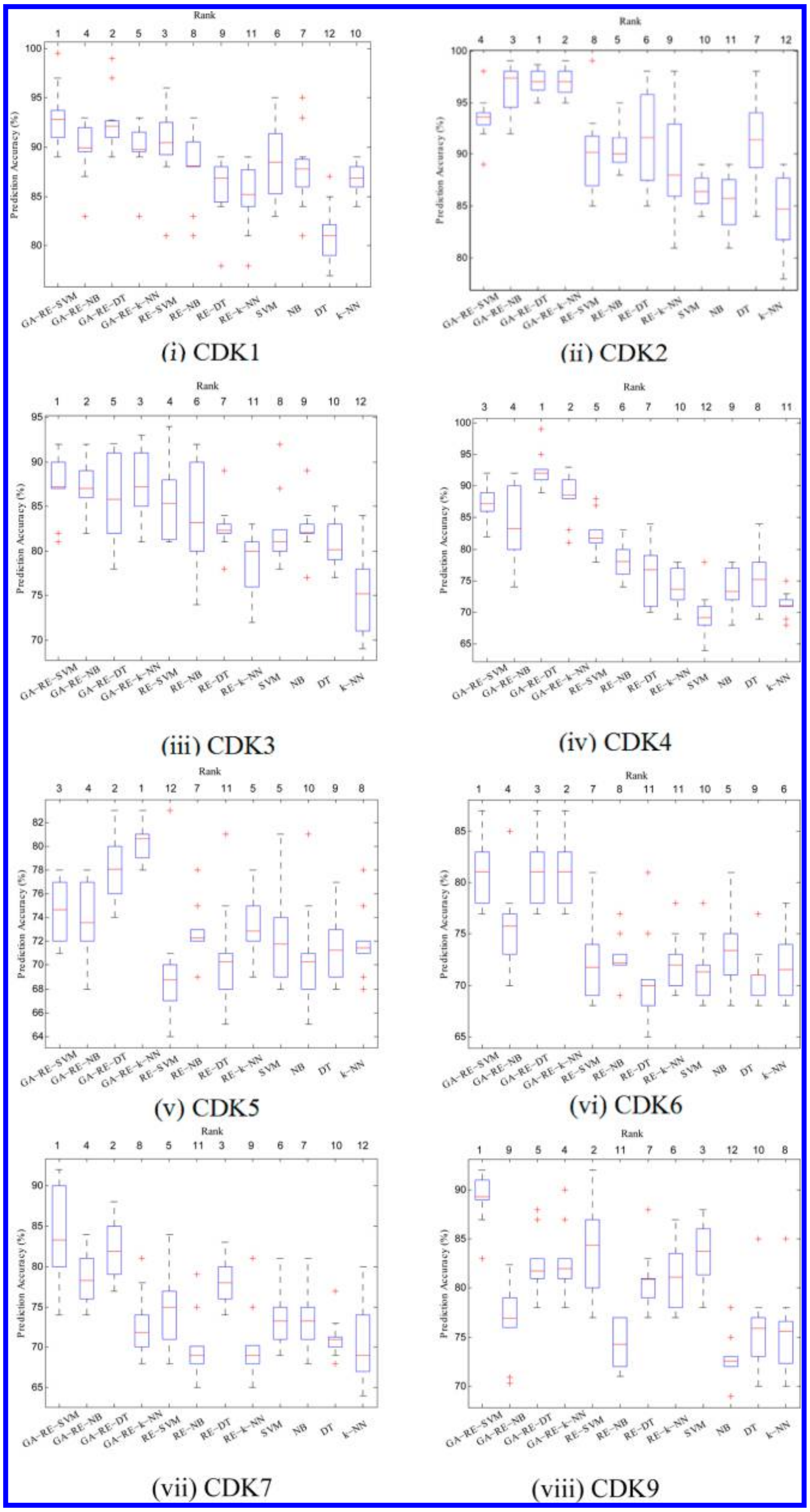
After docking, about six different conformations of each ligand have been obtained from each docking program. Then the VoteDock software is use to select the best conformation for each ligand.[36] VoteDock first collects up to 10 docking poses from each docking program. Thereafter, those poses are compared with each other by computing RMSD matrix between them. For each row of the matrix, the arithmetical mean of RMSD is computed and the pose with lowest value is considered for rescoring phase.

Finally, two scoring functions, X-Score and Prime MMGB-SA, are used to compute binding affinity for chosen best conformations. Results of these rescoring functions provided large number of energy terms, which allow us to select the potential energy terms as features in order to produce better classification accuracy for the new proposed classifier. The whole schematic diagram of your study is shown in Figure 1.

2.2. Analysis of Docking and Rescoring Data. Before the analysis of correlation between experimentally measured IC50 values and docking or rescoring values, IC50 values are normalized in log scale (pIC50) in order to avoid the dominance of higher values. Thereafter, Pearson correlation values are reported in Tables 2 and 3. Pearson correlation value allows us to estimate the strength of our binding affinity prediction (docking and rescoring studies). Note that we have not been considered any compound whose IC50 value lies between 5000 and 10 000 nM. Despite of this fact, still Pearson correlation is found to be a good statistical tool in order to evaluate the quality of binding affinity prediction.

As mentioned before, there have not been a wide docking studies referring to the CDKs. However, Plewczynski et al. made a review where they showed Pearson correlation values varies from 0.1 to 0.38 for FLEXX and eHiTS, respectively, for a large number of different protein families[37] while the recently developed VoteDock reports the experimental reliability up to 0.49 in terms of Pearson correlation value. Our study reports Pearson correlation results varying from 0.0 to 0.48 with an average of 0.19 for docking studies in Table 2 and 0.01 to 0.39 with an average of 0.14 for rescoring studies in Table 3. The ligands for CDKs 3, 4, 6, and 9 produce the higher Pearson correlation values while for other CDKs, it is inferior to the previous results of Plewczynski et al. Therefore, for the future study, it can be worth to select the proper docking program based on higher Pearson correlation value for each CDK, e.g., Glide for CDK2 or eHiTS for CDK5.

The RMSD of 16 ligands and their crystal structures are compared, four examples are shown in Figure 2. The results of RMSD values are in the range of 1.8−3.5 Å with an average of 2.47 Å. This is only less than 0.3−0.5 Å RMSD in compare to VoteDock. It shows that the docking with VoteDock procedure can result in selection of decent conformations. Therefore, it can be stated that the scoring functions are responsible for weak correlations of binding affinities and not for the 3D conformation selection procedure. However, low number of available crystal structures generally does not allow to make this RMSD comparison larger and more statistically reliable. Note that the



(i) CDK1

(ii) CDK2

(iii) CDK3

(iv) CDK4

(v) CDK5

(vi) CDK6

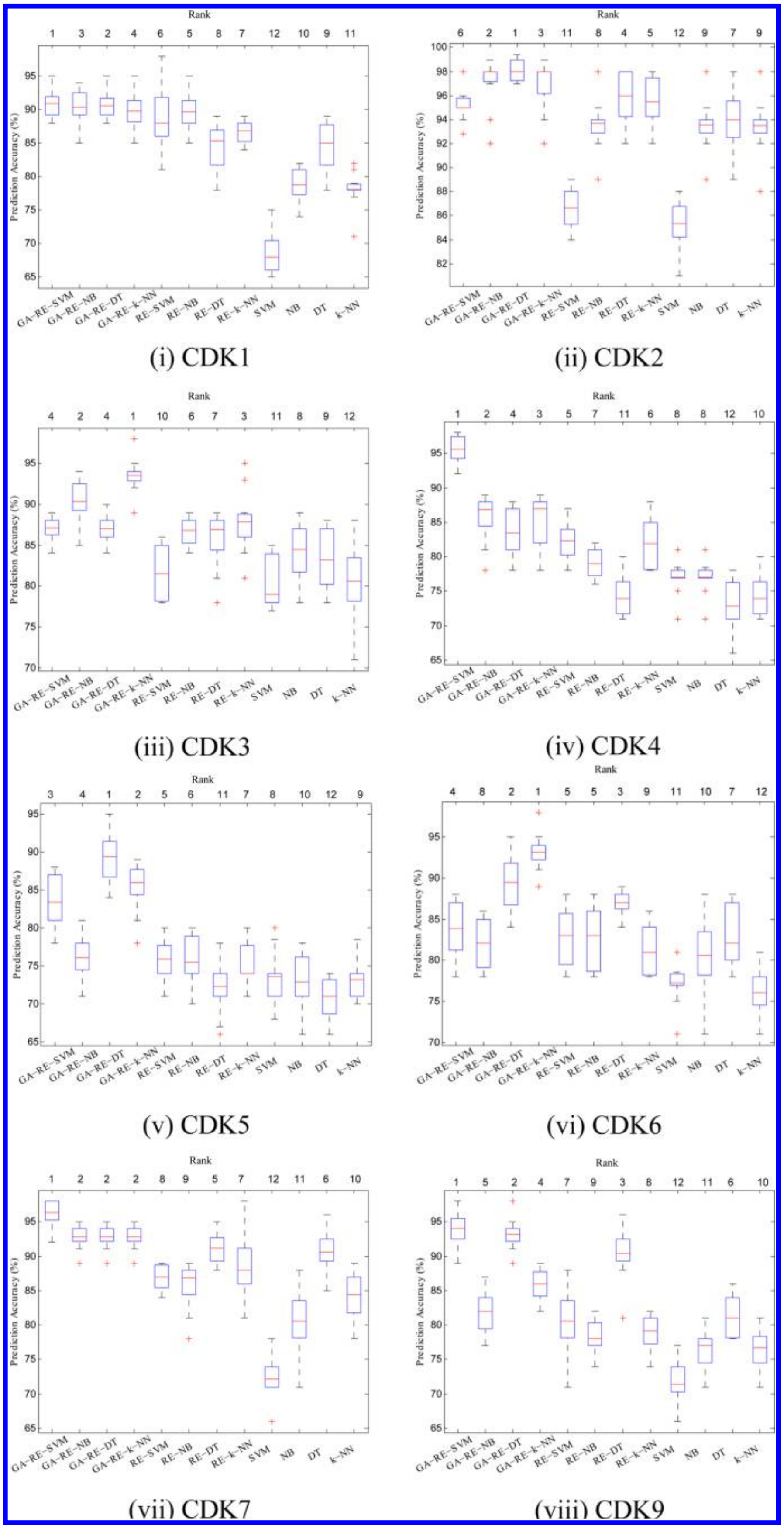
(vii) CDK7

(viii) CDK9

Figure 5. Boxplot of accuracy values of 20 runs for CDK docking data sets.

detailed analysis of CDKs in terms of structural characteristic and physicochemical interactions with its ligands has also been done and mentioned in the Supporting Information.

2.3. Proposed Integrated Classification Technique. The Integrated Classification Technique uses GA,[38−41] RFS[25,26] scheme, and classical Machine Learning methods in the form of ensemble. Here GA is used to select the potential set of features and ensemble size for the underlying ensemble based classification task as an optimization problem. We have used different heterogeneous classification techniques like



(i) CDK1

(ii) CDK2

(iii) CDK3

(iv) CDK4

(v) CDK5

(vi) CDK6

(vii) CDK7

(viii) CDK9

Figure 6. Boxplot of accuracy values of 20 runs for CDK rescoring data sets.

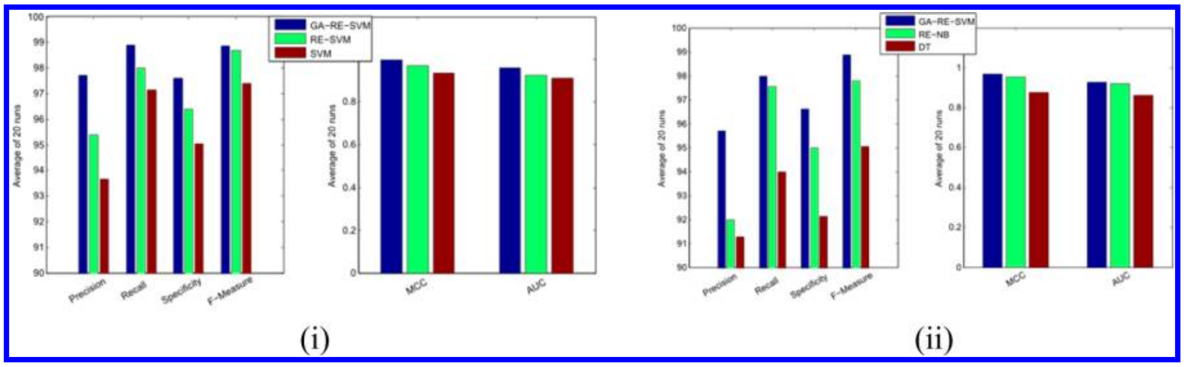


Figure 7. Bar chart representation of average precision, recall, specificity, F-Measure, MCC, and AUC values of the best classification methods at different levels for CDK1 (i) docking and (ii) rescoring data sets.

SVM, NB classifier, DT, and *K*-NN classifier, to produce their ensemble based classification results on the rotation specific feature sets.

In the area of machine learning, many approaches have been developed in the process of appropriate feature selection. However, simultaneous feature selection and ensemble size determination are new in this article. In this regard, GA is used to determine the potential features and ensemble size by optimizing the underlying classification problem. The whole procedure is defined as the level 3 approach while levels 2 and 1 deals with RFS and classification tasks, respectively. This section describes the proposed GA-RE based classification technique in detail along with its block diagrammatic representation in Figure 4.

2.3.1. Representation of Chromosome and Population Initialization. Here the chromosomes contain sequence of binary numbers. A part of this chromosome represents active and inactive features in the data set and other part, computes the ensemble size. For example, for n data points in N dimensions, the length of the chromosome is $N + E$, where the first N positions or genes represent the activity of feature indices and the next E positions represent the maximum number of genes allotted to compute the ensemble size. Here 1 and 0 represent the active and inactive features in the data set respectively and are computed randomly at the initial generation. It is to be noted that a chromosome is valid if the number of 1's in a chromosome for the feature selection part $\in [4, N]$ and the ensemble size $\in [2, 15]$. This process is repeated for each of the I chromosomes in the population, where I is the size of the initial population. Figure 3 illustrates the concept of encoding a chromosome.

2.3.2. Fitness Computation. In fitness computation, the encoded features and ensemble size in each chromosome are decoded. Let $\Phi = \{\mathcal{X}_1, \mathcal{X}_2, .., \mathcal{X}_M\}$ and E^t, where $4 \leq M \leq N$ and $1 \leq t \leq I$, are the set of selected features and ensemble size encoded in a chromosome. Thereafter, fitness is computed using RFS scheme and classifiers ensemble. The RFS scheme creates a diverse set of training points by preparing different nonoverlapping sets of features. To discuss this process, some notations are introduced here. Let us consider a training set consisting of n labeled instances $\mathcal{E} = \{(x_j, y_j)\}_{j=1}^{n}$ in which each instance (x_j, y_j) is described by M input attributes or features and an output attribute, i.e, $x \in \mathbb{R}^n$ and $y \in \mathbb{R}$ where y takes a value from the label space $\{Q_1, Q_2, ... , Q_c\}$. In a classification task, the goal is to use the information only from $\mathcal{E}$ to construct a classifier which can perform well on the unseen data. Let $\mathcal{X}$ be a $n \times M$ matrix consists of M input attributes or features for each training instance and $\mathcal{Y}$ be a one-dimensional column vector contains the output attribute of each training instance in $\mathcal{E}$, therefore, $\mathcal{E}$ can be expressed as concatenating $\mathcal{X}$ and $\mathcal{Y}$ horizontally, i.e., $\mathcal{E} = [\mathcal{X}\ \mathcal{Y}\ ]$. Moreover, the feature set, $\Phi = \{\mathcal{X}_1, \mathcal{X}_2, .., \mathcal{X}_M\}$, splits into F number of feature subsets, where $F \in [2, \lfloor M/2 \rfloor\ ]$. For the ensemble of classifiers, each classifier runs E^t number of times with different rotational feature sets. In order to construct the training data for each of the classifier in ensemble, the following necessary steps are taken.

Step 1: Randomly split Φ into F number of subsets, i.e., F_f^e for simplicity, where e counts the ensemble size and f signifies the current attribute or feature subset. As $F \in [2, \lfloor M/2 \rfloor]$; therefore, the minimum number of subsets is two with at least two features in each subset are considered.

Step 2: Repeat the following steps F times for each subset, i.e., $f = 1, 2, ..., F$.

(a) A new submatrix $\mathcal{X}_f^e$ is constructed which corresponds to the data in matrix $\mathcal{X}$.

(b) From this new submatrix a bootstrap sample $\bar{\mathcal{X}}_f^e$ is drawn where the sample size is generally smaller than that of $\mathcal{X}_f^e$.

(c) Thereafter, $\bar{\mathcal{X}}_f^e$ is used for PCA and the coefficients of all computed principal components are stored into a new matrix ω_f^e.

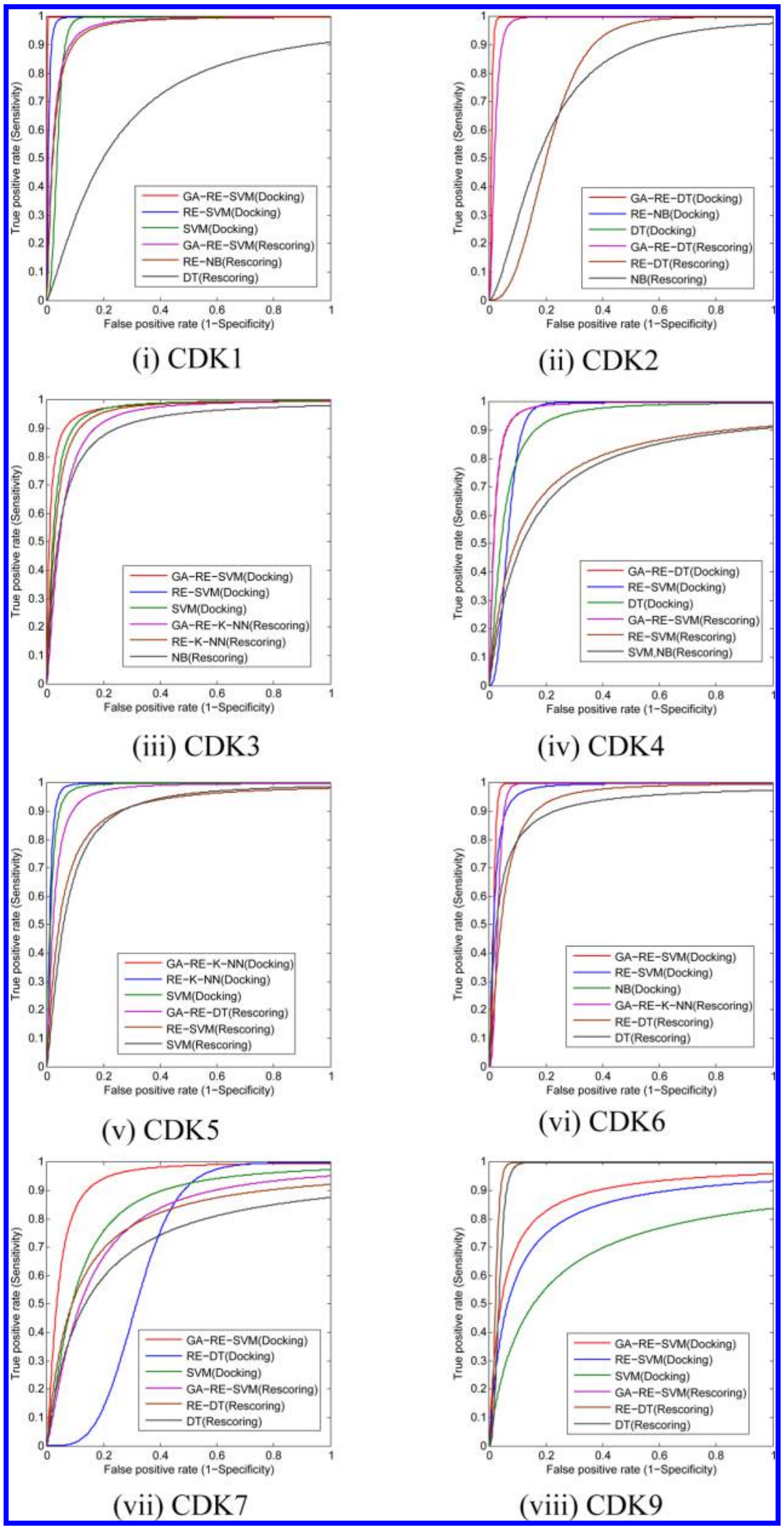
Step 3: In order to have a matrix of size same as the feature set size, arrange each ω_f^e into a block diagonal sparse matrix $\vartheta_f^e\dot{D}$. Once the coefficients in ω_f^e are placed in to the block diagonal sparse matrix $\vartheta_f^e\dot{D}$, the rows of $\vartheta_f^e\dot{D}$ are rearranged so that the order of them corresponds to the original attributes in Φ. During this rearrangement, columns with all zero values are removed from the sparse matrix.

Step 4: The rearranged rotation matrix $\bar{\vartheta}_f^e$ is then used as $[\mathcal{X}\bar{\vartheta}_f^e;\ \mathcal{Y}]$ to create a training data for the classifier. In the classification/testing phase, the test sample, ξ, is multiplied by $\bar{\vartheta}_f^e$ before feeding to the trained classifier. Let $\mathrm{H}_v^e(\xi\bar{\vartheta}_f^e)$ be the posterior probability produced by the classifier H^e on the hypothesis that ξ belongs to class θ_v. Then the confidence for a class is calculated by the average posterior probability of combined base classifiers:

$$\mathcal{Y}_v(\xi) = \frac{1}{E^t}\sum_{e=1}^{E^t}\mathrm{H}_v^e(\xi\bar{\vartheta}_f^e), \qquad v = 1, 2, ..., c$$

Thereafter, ξ is assigned to the class with the largest confidence. However, in this case, we have used 10-fold cross validation in order to evaluate the classification performance of each classifier. Note that all the four steps will repeat for $e = 1, 2, ..., E^t$.

DOI: 10.1021/ci500633c



(i) CDK1

(ii) CDK2

(iii) CDK3

(iv) CDK4

(v) CDK5

(vi) CDK6

(vii) CDK7

(viii) CDK9

Figure 8. ROC plots of best classifiers at different levels for different CDK data sets.

Level 2 classification tasks are shown in Figure 4, which mimics the above steps with the integration of classifiers in order to improve the classification efficacy. Here H represents one of the classifiers, SVM or NB or DT or K-NN, which is used separately for the purpose of classification and validated through 10-fold cross validation. The accuracy is used as an objective function to maximize in order to get the best possible set of features and their different combination of feature sets as a number of subsets (F) as well as ensemble size (E).

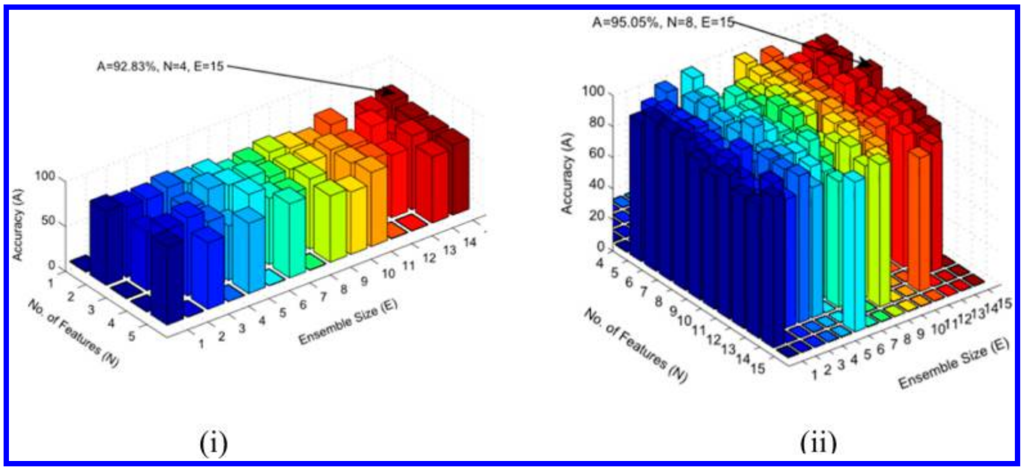


Figure 9. 3D plots of accuracy for the best GA-RE-Classifier out of 20 runs with respect to ensemble size and number of features for CDK4 (i) docking and (ii) rescoring data sets.

2.3.3. Genetic Operators. The tournament selection is used in order to carry the best solutions for the next generation. Thereafter, crossover and mutation are done based on their probability values, Cp and Mp, respectively. A new strategy is applied for the candidate chromosomes to be crossed over. During the crossover, either the odd or even gene/index information of the two crossover chromosomes is exchanged to create new offspring. On the other hand, in the mutation process, each offspring is considered to perturb a single gene by 1 or 0 based on the mutation probability, M_p.

2.3.4. Elitism and Termination Condition. The best solution in the current generation is preserved with the mutant offspring polls for the next generation. The processes of fitness computation, selection, crossover, and mutation are executed for a maximum number of generations. The best chromosome after the final generation will provide the best potential set of features and ensemble size for the classification task. Moreover, the best possible combination of selected features as a subset of feature (F) can be obtained from Level 2.

3. EMPIRICAL RESULTS

The experimental results have been captured for GA-RE-Classifiers, RE-Classifiers, and individual classifiers (ICs) after 20 runs on 16 CDK docking and rescoring data sets separately. In order to avoid the overfitting problem, performance of each classifier has been evaluated using 10-fold cross validation. Note that the list of inactive set of ligands for CDK2 is unavailable in the literature. However, the primary structure of CDK2 protein is 86% homology to CDK1 protein in ATP-binding domain.[43] Thus, the CDK1's negative set of ligands has been considered for CDK2. In GA-RE-Classifiers, GA has been implemented with the parameters like $C_p = 0.8$; $M_p = 0.3$. Moreover, the population size I and maximum generation (gen) have been set to 30 and 100. For the case of classifiers, the parameters of SVM such as kernel function and the soft margin (cost parameter) have been set to 0.5 and 2.0, respectively. Note that an RBF kernel is used here for SVM. In case of K-NN classifier, the best results are obtained while K=7. The GA-RE-Classifiers have been implemented in Matlab version 2012b and provided at http://nucleus3d.cent.uw.edu.pl/cdk/.

The results of GA-RE-Classifiers, RE-Classifiers, and ICs in terms of average accuracy, i.e., the percentage of correctly classified active and inactive binders of CDKs, are reported in Tables 4 and 5 and visually represented by boxplots in Figures 5 and 6. It can be seen from Table 4 that the GA-RE-Classifiers outperformed the other methods for CDK docking data sets. For example, GA-RE-SVM provides the best average accuracy values, 92.47%, 87.00%, 81.11%, 82.56%, and 88.71% for CDK1, CDK3, CDK6, CDK7, and CDK9, while for CDK2 and CDK4, GA-RE-DT provides 98.10% and 92.09% accuracy, and for CDK5, 80.08% accuracy is obtained by GA-RE-KNN. Similarly for rescoring data of CDKs, GA-RE-SVM provides the best accuracy for CDK1, CDK4, CDK7, and CDK9, GA-RE-DT gives the best accuracy for CDK2 and CDK5, while for CDK3 and CDK6, GA-RE-KNN outperforms the others. For each docking and rescoring data, the best classifier at different levels are considered using Tables 4 and 5 in order to compute average precision, recall, specificity, F-Measure, MCC, and AUC values for further analysis. This is reported in Tables S1 and S2 in the Supporting Information. However, the pictorial representation of these metrics is shown in Figure 7 only for CDK1 docking and rescoring data sets, while the others are demonstrated in Figures S4 and S5 in the Supporting Information. From both the tables, the superiority of GA-RE-Classifiers is vividly justified.

ROC curves[42] of the best performing classifiers at different levels either for docking or rescoring data sets are plotted in Figure 8, as this is one of the robust approaches to evaluate the classifiers. The ROC curves show the trade-off between true positive rate (sensitivity) and false positive rate (1 − specificity), while the enclosed region of the curve, i.e., the area under the curve (AUC) gives the value of entire region. Thus, the performance of each classifier is also measured by AUC, which reflects the ability of the predictor in discriminating binders from nonbinders. These curves produced the best AUC values of 0.981 and 0.962 for docking and rescoring data sets that are shown in Tables S1 and S2, respectively, in the Supporting Information. Moreover, in both the tables the average values of precision, recall, specificity, f-measure, and MCC are reported. It is also clear from both the tables that GA-RE-Classifiers provided superior results in terms of AUC and other metrics for all the CDKs.

As the GA-RE-Classifiers are an automatic approach to find the potential features and ensemble size by producing the better classification accuracy, Figure 9 shows the pictorial representation of the best GA-RE-Classifier out of best 20 runs for CDK4 docking and rescoring data sets, while the others are reported in Supporting Information in Figures S6 and S7. It is to be noted that for rescoring data sets of CDK1, CDK4, CDK7, and CDK9, GA-RE-SVM, for CDK2 and CDK5, GA-RE-DT, and for CDK3 and CDK6,

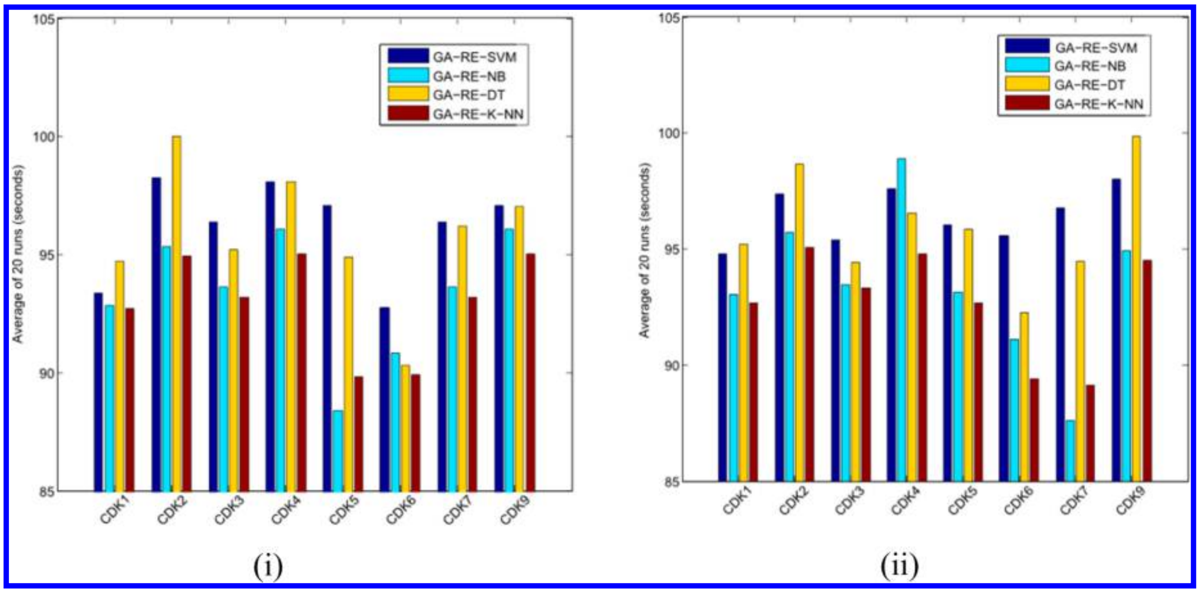


Figure 10. Bar chart representation of CPU time of GA-RE-Classifiers for CDK (i) docking and (ii) rescoring data sets.

GA-RE-KNN classifiers are produced the highest accuracy. The results of these best classifiers in 3D plots are shown in Figures S6 and S7 in Supporting Information for all the CDKs. It is appeared from these figures that the proposed GA-RE-Classifiers produced the better accuracy by finding potential features and ensemble size for accurate prediction of CDK-ligand binding activity.

The computation time of GA-RE-Classifiers is shown in Figure 10(i) and (ii) for CDK docking and rescoring data sets. From the figure, it can be noted that GA-RE-Classifiers took maximum of 100.1S (seconds) to complete the overall execution, whereas GA-RE-SVM which produced the best accuracy for the docking data sets of CDK1, CDK3, CDK6, CDK7, and CDK9, took 93.3S, 97.3S, 92.7S, 96S, and 96.3S, respectively. Therefore, it is clear that the execution time for GA-RE-Classifiers is quite less while getting several benefits like better accuracy in prediction, potential set of features and ensemble size at the same time.

In the last generation of GA-RE-Classifier, best chromosome is obtained that encodes the potential features and ensemble size. Moreover, the information of the subset numbers of the selected features is kept in another variable. For each GA-RE-Classifier, based on the best of 20 accuracy values, in Table 6, ensemble size, number of subsets, and number of selected features are reported for CDK rescoring data sets, while the same for CDK docking data sets, is reported in Table S3 in the Supporting Information. Along with these results, Tables S3 and S4 in the Supporting Information, also contain the name of the selected features. For example, GA-RE-SVM took eight features out of 17 and ensemble of 15 SVMs to produce the best accuracy of 95.05% for CDK4 rescoring data set.

3.1. Statistical Analysis of Results.

The statistical significance of all the classifiers results is analyzed by using the Friedman test.[33,34] Generally, the Friedman test ranks the classifiers for each data set separately. To compute the average rank R_j, let r_i^j be the rank of the jth algorithm for the ith data set where the number of data sets and algorithms are U and Q, respectively. Therefore, the average rank is

$$R_j = \frac{1}{U} \sum_i r_i^j$$

Table 6. Statistic of Number of Original Features, Ensemble Size, Number of Feature Subsets and Number of Selected Features Produced by GA-RE-Classifiers for Rescoring Data Sets

Data Set	GA-RE-Classifier	number of original features	ensemble size	number of feature subsets	number of selected features
CDK1	SVM	17	6	3	11
	NB		8	5	13
	DT		1	3	10
	K-NN		15	2	8
CDK2	SVM		9	6	12
	NB		4	2	6
	DT		13	2	4
	K-NN		14	4	10
CDK3	SVM		9	3	13
	NB		9	6	12
	DT		13	2	4
	K-NN		14	3	8
CDK4	SVM		15	3	8
	NB		9	6	12
	DT		1	3	10
	K-NN		15	2	8
CDK5	SVM		1	3	7
	NB		14	4	10
	DT		1	2	9
	K-NN		14	4	10
CDK6	SVM		8	4	8
	NB		12	3	12
	DT		1	3	7
	K-NN		1	3	7
CDK7	SVM		9	6	12
	NB		14	3	8
	DT		1	3	10
	K-NN		2	3	8
CDK9	SVM		8	4	8
	NB		13	2	4
	DT		9	3	9
	K-NN		15	2	8

Table 7. Friedman Ranks of All Classifiers for CDK Docking Data Sets

	GA-RE-Classifier (Level 3)				RE-Classifier (Level 2)				Classifier (Level 1)			
data set	SVM	NB	DT	*K*-NN	SVM	NB	DT	*K*-NN	SVM	NB	DT	*K*-NN
CDK1	1	4	2	5	3	8	9	11	6	7	12	10
CDK2	4	3	1	2	8	5	6	9	10	11	7	12
CDK3	1	2	5	3	4	6	7	11	8	9	10	12
CDK4	3	4	1	2	5	6	7	10	12	9	8	11
CDK5	3	4	2	1	12	7	11	5.5	5.5	10	9	8
CDK6	1	4	3	2	7	8	11.5	11.5	10	5	9	6
CDK7	1	4	2	8	5	11	3	9	6	7	10	12
CDK9	1	9	5	4	2	11	7	6	3	12	10	8
average rank	**1.87**	**4.25**	**2.62**	**3.37**	**5.75**	**7.75**	**7.69**	**9.12**	**7.56**	**8.75**	**9.37**	**9.87**

Table 8. Friedman Ranks of All Classifiers for CDK Rescoring Data Sets

	GA-RE-Classifier (Level 3)				RE-Classifier (Level 2)				Classifier (Level 1)			
data set	SVM	NB	DT	*K*-NN	SVM	NB	DT	*K*-NN	SVM	NB	DT	*K*-NN
CDK1	1	3	2	4	6	5	8	7	12	10	9	11
CDK2	6	2	1	3	11	8	4	5	12	9.5	7	9.5
CDK3	4.5	2	4.5	1	10	6	7	3	11	8	9	12
CDK4	1	2	4	3	5	7	11	6	8.5	8.5	12	10
CDK5	3	4	1	2	5	6	11	7	8	10	12	9
CDK6	4	8	2	1	5.5	5.5	3	9	11	10	7	12
CDK7	1	3	3	3	8	9	5	7	12	11	6	10
CDK9	1	5	2	4	7	9	3	8	12	11	6	10
average rank	**2.69**	**3.62**	**2.43**	**2.62**	**7.19**	**6.94**	**6.5**	**6.5**	**10.80**	**9.75**	**8.5**	**10.43**

In this test, under the null hypothesis, all the algorithms are equivalent and so their ranks R_j should be equal. The Friedman statistic (chi-square value) is computed as follows.

$$H_F = \frac{12}{UQ(Q+1)} \sum_{J}^{Q} R_j^{\,2} - 3U(Q-1)$$

The Friedman statistic is distributed according to H_F with $Q - 1$ degrees of freedom, when $U > 10$ and $Q > 5$. For a smaller number of algorithms and data sets, exact critical values are computed.

Tables 7 and 8 report the ranks of all the classifiers for 16 CDK data sets of docking and rescoring. Moreover, using average ranks it can be stated that GE-RE-SVM is the most promising among others. For both the cases, it achieves the ranks 1.87 and 2.69. Also the average ranks are used to compute the H_F values, i.e., 51.983 and 62.892 for docking and rescoring data sets, respectively. Therefore, the corresponding p-values are 0.199×10^{-4} and 0.301×10^{-4} at $\alpha = 0.05$ significance level, which also emphasize the acceptance of the results produced by the methods, is statistically significant for docking and rescoring data sets of CDKs.

4. CONCLUSION

Understanding the protein−ligand binding interaction is an essential task for designing more selective and potent inhibitors. In this study, we have developed a new classification technique that confirms the binding activity of CDK-ligands with high accuracy after analyzing molecular docking and rescoring data of known CDK-ligands binders. The developed technique is called Genetic Algorithm integrated Rotational Ensemble based Classifier, where a Rotational Feature Selection scheme is used with the traditional classifiers like Support Vector Machine, Naive Bayesian classifier, Decision Tree, and *K*-Nearest Neighbor classifiers separately in ensemble form. The developed technique can automatically find the potential features and ensemble size by achieving the higher accuracy for 16 docking and rescoring data sets of CDKs. The experimental results show that the developed technique outperformed the traditional classification approaches. This is also statically validated through a Friedman test.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

structural analysis of CDKs, definitions of various metrics, and suporting figures and tables. The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/ci500633c.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: d.plewczynski@cent.uw.edu.pl.

Author Contributions

○I.S., B.R., and S.S.B. contributed equally.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was partially supported by the Polish National Science Centre (Grant number UMO-2013/09/B/NZ2/00121 and 2014/15/B/ST6/05082), COST BM1405 and BM1408 EU actions, European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no.: 246016, and the Erasmus Mundus Mobility with Asia (EMMA) grant. Moreover, we are thankful to the reviewers for their critical comments and suggestions in order to improve this article.

REFERENCES

(1) Srinivasan, P.; Manikandan, R.; Arulvasu, C. Inhibition of Cyclin Dependent Kinase-2 and Glycogen Synthase Kinase-3 by Herbal Derivative 12-disubstituted Idopyranose Through in Silico Analysis. *Journal of Advanced Scientific Research.* **2012**, *3* (1), 65−72.

(2) Morgan, D. Cyclin-Dependent Kinases: Engines, Clocks, and Microprocessors. *Annu. Rev. Cell Dev. Biol.* **1997**, *13* (1), 261−291.

(3) Schwartz, G. K.; Shah, M. A. Targeting the Cell Cycle: A New Approach to Cancer Therapy. *J. Clin. Oncol.* **2005**, *23* (36), 9408−9421.

(4) Malumbres, M.; Barbacid, M. Cell Cycle Kinases in Cancer. *Curr. Opin. Genet. Dev.* **2007**, *17* (1), 60−65.

(5) Morris, G. M.; Lim-Wilby, M. Molecular Docking.*Molecular Modeling of Protiens*; Methods Molcular Biology; Springer, 2008; Vol. *443* (1), pp 365−382.

(6) Jain, A. N. Surflex: Fully Automatic Flexible Molecular Docking Using a Molecular Similarity-Based Search Engine. *J. Med. Chem.* **2003**, *46*, 499−511.

(7) Pagani, I.; Liolios, K.; Jansson, J.; Chen, I. M.; Smirnova, T.; Nosrat, B.; Markowitz, V. M.; Kyrpides, N. C. The Genomes Online Database (Gold) V.4: Status of Genomic and Metagenomic Projects and Their Associated Metadata. *Nucleic Acids Res.* **2012**, *40*, D571−9.

(8) Zsoldos, Z.; Reid, D.; Simon, A.; Sadjad, B. S.; Johnson, A. P. Ehits: An Innovative Approach to the Docking and Scoring Function Problems. *Current Protein and Peptide Science.* **2016**, *7*, 421−35.

(9) Wu, G.; Robertson, D. H.; Brooks, C. L.; Vieth, M. Detailed Analysis of Grid-Based Molecular Docking: A Case Study of Cdocker-a Charmm-Based Md Docking Algorithm. *J. Comput. Chem.* **2003**, *24*, 1549−1562.

(10) Rao, S. N.; Head, M. S.; Kulkarni, A.; LaLonde, J. M. Validation Studies of the Site-Directed Docking Program Libdock. *J. Chem. Inf. Model.* **2007**, *47*, 2159−2171.

(11) Friesner, R.; Banks, J.; Murphy, R.; Halgren, T.; Klicic, J.; Mainz, D.; Repasky, M.; Knoll, E.; Shelley, M.; Perry, J.; Shaw, D.; Francis, P.; Shenkin, P. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47* (7), 1739−1749.

(12) Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R. Principles of Docking: An Overview of Search Algorithms and A Guide to Scoring Functions. *Proteins: Struct., Funct., Genet.* **2002**, *47* (4), 409−443.

(13) Zhang, X.; Wang, T.; Luo, H.; Yang, J. Y.; Deng, Y.; Tang, J.; Yang, M. Q. 3d Protein Structure Prediction with Genetic Tabu Search Algorithm. *BMC Syst. Biol.* **2010**, *4* (Suppl 1), S6.

(14) Sauton, N.; Lagorce, D.; Villoutreix, B. O.; Miteva, M. A. MS-DOCK: Accurate Multiple Conformation Generator and Rigid Docking Protocol for Multi-Step Virtual Ligand Screening. *BMC Bioinf.* **2008**, *9*, 184.

(15) Taylor, R. D.; Jewsbury, P. J.; Essex, J. W. A Review of Protein-Small Molecule Docking Methods. *J. Comput.-Aided Mol. Des.* **2002**, *16* (3), 151−166.

(16) Holt, P. A.; Chaires, J. B.; Trent, J. O. Molecular Docking of Intercalators and Groove-Binders to Nucleic Acids Using Autodock and Surflex. *J. Chem. Inf. Model.* **2008**, *48*, 1602−1615.

(17) DesJarlais, R. L.; Sheridan, R. P.; Dixon, J. S.; Kuntz, I. D.; Venkataraghavan, R. Docking Flexible Ligands to Macromolecular Receptors by Molecular Shape. *J. Med. Chem.* **1986**, *29* (11), 2149−2153.

(18) Huang, N.; Kalyanaraman, C.; Irwin, J. J.; Jacobson, M. P. Hysics-based Scoring of Protein-Ligand Complexes: Enrichment of Known Inhibitors in Large-Scale Virtual Screening. *J. Chem. Inf. Model.* **2006**, *46* (1), 243−253.

(19) Huang, S. Y.; Grinter, S. Z.; Zou, X. Scoring Functions and Their Evaluation Methods for Protein-Ligand Docking: Recent advances and future directions. *Phys. Chem. Chem. Phys.* **2010**, *12* (1), 12899−12908.

(20) Muegge, I. A Knowledge-based Scoring Function for Protein-Ligand Interactions: Probing the Reference State. *Virtual Screening: An Alternative or Complement to High Throughput Screening*; Springer, 2002; Vol. *20*, pp 99−114.

(21) Zhong, S.; Zhang, Y.; Xiu, Z. Rescoring Ligand Docking Poses. *Curr. Opin. Drug Discov. Dev.* **2010**, *13*, 326−334.

(22) Wang, R.; Lai, L.; Wang, S. Further Development and Validation of Empirical Scoring Functions for Structure-based Binding Affinity Prediction. *J. Comput.-Aided Mol. Des.* **2002**, *16* (1), 11−26.

(23) Mulakala, C.; Viswanadhan, V. N. Could MM-GBSA be Accurate Enough for Calculation of Absolute Protein/Ligand Binding Free Energies? *J. Mol. Graphics Modell.* **2013**, *46*, 41−51.

(24) Ghosh, A.; Rapp, C. S.; Friesner, R. A. Generalized Born Model based on a Surface Integral Formulation. *J. Phys. Chem. B* **1998**, *102* (52), 10983−10990.

(25) Bhowmick, S. S.; Saha, I.; Rato, L.; Bhattacharjee, D. RotaSVM: A New Ensemble Classifier. *EVOLVE - A Bridge between Probability, Set Oriented Numerics, and Evolutionary Computation IV*; Advances in Intelligent Systems and Computing; Springer, 2013; Vol. *227*, pp 47−57.

(26) Bhowmick, S. S.; Saha, I.; Rato, L.; Bhattacharjee, D. Improving Performance of Classifiers using Rotational Feature Selection Scheme. In *Proceedings of the 2nd International Conference on Advances in Computer Science and Engineering (CSE 2013)*, Pittsburgh, PA, July 27−29th, 2013; pp 309−314.

(27) Boser, B. E.; Guyon, I. M.; Vapnik, V. N.A Training Algorithm for Optimal Margin Classifiers. In *Proceedings of the 5th Annual Workshop on Computational Learning Theory*, Pittsburgh, PA, July 27−29, 1992; pp 144−152.

(28) Duda, R. O.; Hart, P. E. *Pattern Classification and Scene Analysis*; John Wiley & Sons: New York, 1973.

(29) Quinlan, J. R. *C4.5: Programs for Machine Learning*; Morgan Kaufmann Publishers: San Francisco, CA, 1993.

(30) Sun, S. L. Ensembles of Feature Subspaces for Object Detection. *Advances in Neural Networks−ISSN 2009*; Lecture Notes in Computer Science; Springer, 2009; Vol. *5552*, pp 996−1004.

(31) Cover, T.; Hart, P. Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory.* **1967**, *13* (1), 21−27.

(32) Hotelling, H. Analysis of a Complex of Statistical Variables into Principal Components. *Journal of Educational Psychology.* **1933**, *24*, 417−441.

(33) Friedman, M. The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *J. Am. Stat. Assoc.* **1937**, *32*, 675−701.

(34) Friedman, M. A Comparison of Alternative Tests of Significance for the Problem of m Rankings. *Ann. Math. Stat.* **1940**, *11*, 86−92.

(35) Roy, A.; Kucukural, A.; Zhang, Y. I-TASSER: A Unified Platform for Automated Protein Structure and Function Prediction. *Nat. Protoc.* **2010**, *5* (4), 725−738.

(36) Plewczynski, D.; Lazniewski, M.; von Grotthuss, M.; Rychlewski, L.; Ginalski, K. VoteDock: Consensus Docking Method for Prediction of Protein-Ligand Interactions. *J. Comput. Chem.* **2011**, *32* (4), 568−581.

(37) Plewczynski, D.; Lazniewski, M.; Augustyniak, R.; Ginalski, K. Can We Trust Docking Results? Evaluation of Seven Commonly Used Programs on PDBbind Database. *J. Comput. Chem.* **2011**, *32* (4), 742−755.

(38) Goldberg, D. E. Genetic Algorithms in Search, *Optimization and Machine Learning*; Addison-Wesley: New York, 1989.

(39) Davis, L. *Handbook of Genetic Algorithms*; Van Nostrand Reinhold: New York, 1991.

(40) Michalewicz, Z. *Genetic Algorithms + Data Structures = Evolution Programs*; Springer: New York, 1992.

(41) Filho, J. L. R.; Treleaven, P. C.; Alippi, C. Genetic Algorithm Programming Environments. *Computer* **1994**, *27* (6), 28−44.

(42) Swets, J. A. Measuring the Accuracy of Diagnostic Systems. *Science* **1988**, *240* (4857), 1285−1293.

(43) Vassilev, L.; Tovar, C.; Chen, S.; Knezevic, D.; Zhao, X.; Sun, H.; Heimbrook, D.; Chen, L. Selective Small-Molecule Inhibitor Reveals Critical Mitotic Functions of Human CDK1. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103* (28), 10660−10665.

(44) Khuntawee, W.; Rungrotmongkol, T.; Hannongbua, S. Molecular Dynamic Behavior and Binding Affinity of Flavonoid Analogues to the Cyclin Dependent Kinase 6/cyclin D Complex. *J. Chem. Inf. Model.* **2012**, *52* (1), 76−83.

(45) Schwartz, G. K.; Shah, M. A. Targeting the Cell Cycle: A New Approach to Cancer Therapy. *J. Clin. Oncol.* **2005**, *23* (36), 9408−9421.

(46) Bolton, E. E.; Wang, Y.; Thiessen, P. A.; Bryant, H. B. PubChem: Integrated Platform of Small Molecules and Biological Activities. *Annual Reports in Computational Chemistry*; Elsevier, 2008; Vol. *4*, pp 217−241.
(47) Liu, T.; Li, Y.; Wen, X.; Jorissen, N. R.; Gilson, K. M. BindingDB: A Web-Accessible Database of Experimentally Determined Protein-Ligand Binding Affinities. *Nucleic Acids Res.* **2007**, *35*, D198−D201.
(48) Kurapati, R. K.; Giri, A.; Nadendla, R. R. Cross Docking as a Method to Select CDK-9 Protein Target for Virtual Screening Studies. *International Journal of Computational Bioinformatics and In Silico Modeling* **2013**, 2 (6), 275−277.
(49) Akrimah, D. H. T.; Musadad, A. Docking of Potent Anticancer Agents; 4-(Pyrazol-4yl) Pyrimidine Derivatives as Selective Cyclin Dependent Kinase 4/6 Inhibitors. *Int. J. Chem. Eng. Appl.* **2010**, *4* (6), 419−422.

■ NOTE ADDED AFTER ASAP PUBLICATION

There were errors in Figures 5, 6, and 8 in the version published ASAP July 10, 2015; the corrected version was published ASAP July 27, 2015.