RAIRO-Oper. Res. 50 (2016) 387–400
DOI: 10.1051/ro/2015042

RAIRO Operations Research
www.rairo-ro.org

MILP-HYPERBOX CLASSIFICATION FOR STRUCTURE-BASED DRUG DESIGN IN THE DISCOVERY OF SMALL MOLECULE INHIBITORS OF SIRTUIN6

Mehmet Tardu[1], Fatih Rahim[2], I. Halil Kavakli[3,4] and Metin Turkay[2]

Abstract. Virtual screening of chemical libraries following experimental assays of drug candidates is a common procedure in structure-based drug discovery. However, virtual screening of chemical libraries with millions of compounds requires a lot of time for computing and data analysis. *A priori* classification of compounds in the libraries as low- and high-binding free energy sets decreases the number of compounds for virtual screening experiments. This classification also reduces the required computational time and resources. Data analysis is demanding since a compound can be described by more than one thousand attributes that make any data analysis very challenging. In this paper, we use the hyperbox classification method in combination with partial least squares regression to determine the most relevant molecular descriptors of the drug molecules for an efficient classification. The effectiveness of the approach is illustrated on a target protein, SIRT6. The results indicate that the proposed approach outperforms other approaches reported in the literature with 83.55% accuracy using six common molecular descriptors (SC-5, SP-6, SHBd, minHaaCH, maxwHBa, FMF). Additionally, the top 10 hit compounds are determined and reported as the candidate inhibitors of SIRT6 for which no inhibitors have so far been reported in the literature.

Mathematics Subject Classification. 90C11.

Received September 8, 2015. Accepted September 21, 2015.

1. Introduction

Drugs are necessary for the prevention and treatment of diseases. Many diseases such as cancer, diabetes, and Alzheimer constantly threaten the quality of life of humans. Although new and effective drugs are always in high demand, the ability to design and discover them has not been very successful over the years due to the lack of systematic approaches that can integrate different aspects of the drug discovery process. The process of drug development is challenging, time consuming, expensive, and requires consideration of different aspects such as effectiveness, side effects, and toxic effects. Combinatorial chemistry has emerged as a new paradigm in the field of drug discovery as it provides a large library of compounds at a time. Moreover, the number of commercially available compounds are also continuously increasing. So the process of compound selection and prioritization

Keywords. Structure-based drug design, SIRT6, MILP-HB.

[1] Department of Computational Science and Engineering, Koc University, 34450 Istanbul, Turkey.
[2] Department of Industrial Engineering, Koc University, 34450 Istanbul, Turkey. mturkay@ku.edu.tr
[3] Department of Molecular Biology and Genetics, Koc University, 34450 Istanbul, Turkey.
[4] Department of Chemical and Biological Engineering, Koc University, 34450 Istanbul, Turkey.

Article published by EDP Sciences

is very crucial to reduce the time and computational cost of screening the libraries of compounds [23,29,39]. To fulfill these challenges, several multidisciplinary approaches are required in the process of drug development.

Structure-based drug design methods are an integral part of drug development for known 3D structures of potential drug binding sites, which are the active sites. In structure-based drug design, a known 3D structure of a target bound to its natural ligand is determined either by X-ray crystallography or by NMR to identify its binding site, also called the active site. For the discovery of drug candidates for a known target, this is the starting point of structure-based drug design. Once the ligand bound 3D structure is known, a virtual screening of large collections of chemical compounds from different databases, such as ZINC [16], can be performed. Virtual screening provides a score based on the steric and electrostatic interactions of the drug candidates with the target protein. The scoring function provides a computational estimate such as binding free energy (BFE) and binding constant for the activity of the drug candidate on the target site. Next, selected drug candidates with relatively high binding affinity (low BFE) are tested with *in vitro* and *in vivo* assays [17, 22]. However, virtual screening of those small molecule libraries with millions of compounds requires considerable computing and data analysis time. To reduce computational time and resources during the virtual screening, we have combined a machine learning and virtual screening approach.

Our motivation behind this study is to filter the initial small molecule libraries that contain compounds that may be deemed unsuitable for screening against a particular target protein by using a classification algorithm, mixed-integer linear programming based hyperbox (MILP-HB) [38]. The objective is to screen out the compounds in the initial library that have high-BFE and to select low-BFE compounds that are specific to the target protein. For this purpose, we targeted SIRT6 protein to find novel inhibitor (low-BFE) compounds. The biological importance of SIRT6 protein is discussed in the next section in detail. In our approach, to classify compounds as low- and high-BFE sets, we used the partial least square regression (PLSR) combined with MILP-HB method that takes the molecular descriptors as attributes of the model.

PaDEL-Descriptor [43] software was used to calculate the molecular descriptor values of the compounds in the small molecule library. The software provided 1975 descriptors in 12 categories of fingerprints for each compound. Since the number of descriptors were excessively high, most of them were expected to be irrelevant for the prediction of BFE values for the specific target protein. It was, therefore, important to select a subset of descriptors that were most informative about BFE values between the compounds and target protein. The molecular descriptor set was reduced by using the Unsupervised Forward Selection (UFS) method [40]. Then, a PLSR model [42] was constructed to eliminate the redundant set of descriptors for the prediction of BFE values. The purpose of using a PLSR model was to find the most informative molecular descriptors to be used in building the MILP-HB classifier. After conducting regression analysis, selecting the relevant molecular descriptors for MILP-HB, and making preliminary classification studies, significance tests were performed on the descriptors selected by the regression method to improve the classification accuracies. The strength of the MILP-HB algorithm not only comes from combining regression with classification but also the ability to improve the classification accuracies by its iterative approach. Previous studies showed that its classification efficiency and accuracy is superior to alternative methods [1,7,8,18].

The comparison of MILP-HB classification accuracy with the accuracies of the classification methods available in the WEKA data mining package [11] was also made. WEKA is a collection of machine learning algorithms for data mining tasks. It contains 63 different classification methods, but in this paper, we reported only 12 methods that provided the highest accuracy to save space. A brief overview of these classifiers was presented in the 'Strategies and Methods' section. Comparison of classification accuracies between MILP-HB and WEKA classifiers were reported in the 'Data and Results' section. Our approach outperformed all of the classifiers available in WEKA with 83.55% accuracy. We concluded that the 6 most significant descriptors (SC-5, SP-6, SHBd, minHaaCH, maxwHBa, FMF) that we selected by using PLSR provide the highest classification accuracy with MILP-HB method. Then, 4 million compounds in the small molecule library were classified according to those 6 molecular descriptors as low- and high-BFE sets by using MILP-HB. About 2.4 million compounds were classified as low-BFE compounds that were further subjected to virtual screening by docking to SIRT6 protein. After molecular docking, the compounds with minimum BFEs were selected and analyzed according

to important characteristics such as docking positions, hydrogen bonding interactions with the residues at the NAD+ binding site and pocket occupancy. The top 10 hit compounds were determined and reported as the candidate inhibitors of SIRT6. BFEs of the selected candidates were in the range of −11.2 to −11.9 kcal/mol.

2. The biological importance of sirtuin6

Sirtuins belong to the Class III family of histone deacetylase enzymes (HDACs) that require the cofactor NAD+ as a substrate [10, 13]. Sirtuins catalyze the removal of an acetyl moiety from the ϵ-amino group of lysine residues within protein targets to yield the deacetylated protein product, nicotinamide, and 2'-O-acetyl-ADP-ribose. Sirtuins regulate various biological processes, such as cell survival, transcription, metabolism, DNA damage/repair, and longevity by deacylating different proteins, including histones, transcription factors, and metabolic enzymes [10].

There are seven mammalian Sirtuins (SIRT1-7) that show different intracellular localization and deacetylate different sets of substrate proteins. Among the Sirtuins, SIRT6 is a key enzyme in multiple molecular pathways related to aging, DNA repair, telomere maintenance, glycolysis and cancer [13]. SIRT6, which is localized in the nucleus, displays deacetylase and weak ADP-ribosyl transferase activity [13, 25]. Within the nucleus it is associated with heterochromatic regions [28] and modifies two key enzymes involved in double-strand break repair, PARP1 [19] and CtIP [27]. Also, SIRT6 is shown to be involved in Base Excision Repair (BER) and its loss leads to age-associated degenerative processes [30].

Recent studies showed that SIRT6 deficiency promotes tumor-genesis by promoting increased glycolysis by regulating the expression of glycolytic genes, including the glucose transporter GLUT1, in a HIF1$\propto$-dependent manner [44]. In addition, p53 positively regulates the transcriptional levels of SIRT6 under normal conditions in tumor cells [20]. Moreover, SIRT6 has been shown to interact with the relA subunit of NFB and deacetylate histone H3 lys9 at the NFB target gene promoters resulting in silencing gene transcription involved in apoptosis and cellular senescence [21].

It is obvious that regulating the activity of the SIRT6 with small molecule inhibitors will enable us to use molecules in the treatment of metabolic disorders, such as diabetes (as blood glucose-lowering agents), in aging related degenerative processes, and in cancer (as chemosensitizers) [9]. Additionally, such molecules will help us to study SIRT6 dependent biological processes and pathways *in vitro* and *in vivo*. In this context, availability of the SIRT6 crystal structure (pdb id: 3K35) represents an important tool for structure-based in-silico screenings and for general understanding of the mode of action of SIRT6 and of its chemical modulators.

3. Strategies and methods

Strategies followed in this study are summarized in Figure 1. First, the Molecular Dynamics (MD) simulation was performed on SIRT6 protein. This step was applied to calculate binding energies at realistic conditions that are close to physiological conditions. Then, a small molecule library was prepared by using several databases. Library size was decreased from 7 million to 4 million compounds by applying Lipinski's rule of five criteria [24]. After that, virtual screening was performed on a subset of randomly selected compounds to calculate BFE values.

A data mining study was conducted to classify the compounds as low- and high-BFE classes using the MILP-HP method.BFE of NAD+ (−8.3 kcal/mol), which is a substrate of SIRT6, was used as a threshold to classify molecules as low-BFE (above −8.3 kcal/mol) or high-BFE (below −8.3 kcal/mol) sets. Molecular descriptors of the compounds were generated by using PaDEL software. The descriptor set was reduced first by UFS and then by the PLSR model. The 15 descriptors with the highest regression coefficients were added to the set of 15 most significant descriptors, and the first 6 of them were used as input for the initial classification by MILP-HB. The accuracy of the MILP-HB classification was compared with other classifiers in the WEKA data mining package. Then, this approach with 6 specific molecular descriptors, SC-5, SP-6, SHBd, minHaaCH, maxwHBa and FMF, helped us to reduce the size of library from 4 million to 2.4 million compounds. The remaining 2.4 million

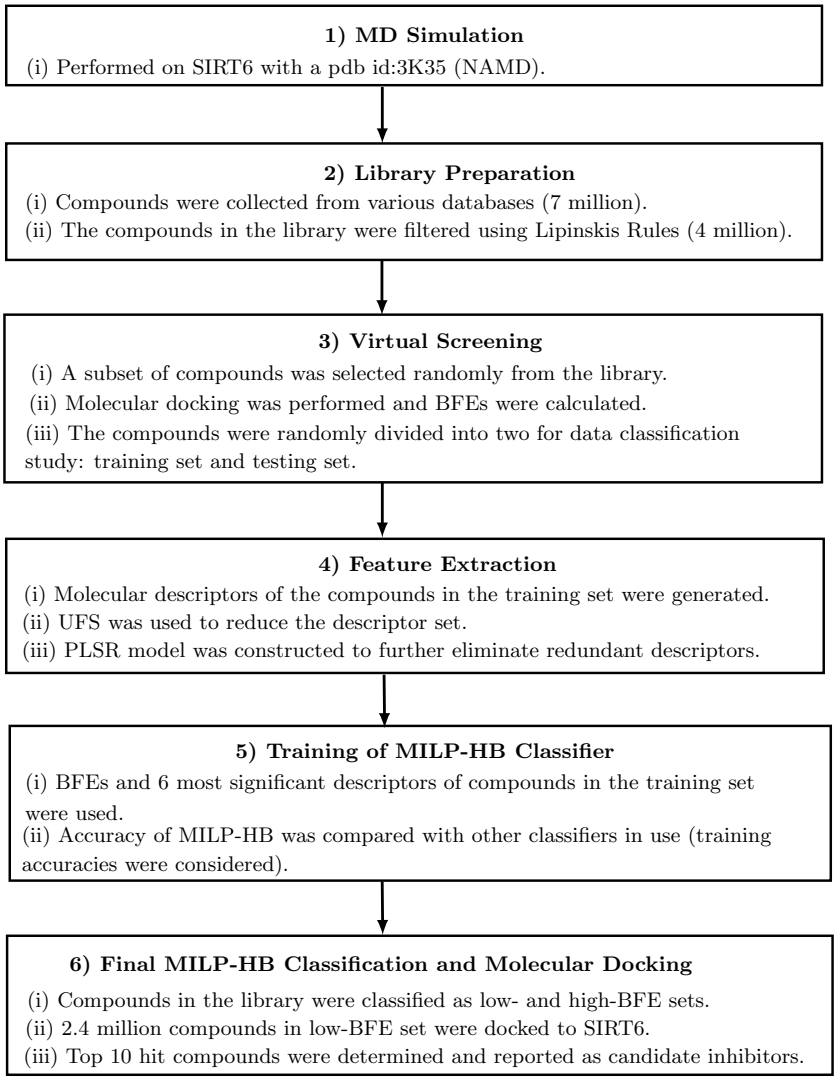


FIGURE 1. Systematic approach used in the study.

compounds in the low-BFE set were docked to SIRT6. Finally, the compounds with minimum BFEs were selected and analyzed and, finally, the top 10 hit compounds were determined and reported as the candidate inhibitors of SIRT6.

3.1. Molecular dynamics simulation

The reported structure of human SIRT6 was in a complex with adenosine-5-diphosphoribose at 2.0 Å [31]. MD simulations were carried out by using the NAMD software, version 2.6 with the PARAM22 version of the CHARMM force field [26, 32]. The protein was solvated in a rectangular box including TIP3P water molecules and counter ions. First, the system was minimized through 10 000 steps by keeping the backbone fixed, and then backbone atoms were relaxed through 10 000 steps. Next, the system was heated to 310 K with 10 K increments (10 ps simulation at each increment). After the equilibration of the system, a MD simulation was carried out with constant temperature (310 K) and pressure control using the Langevin piston method. The time-step of the simulation was set to be 2 fs and the bonded interactions, the van der Waals interactions (12 Å cut-off),

and the long-range electrostatic interactions with particle-mesh Ewald (PME) were included in the calculations to define the forces acting on the system. The distances between the atoms of the molecules are calculated. A survey of the algorithms can be found in [36]. The damping coefficient was set to be 5 ps^{-1} using Langevin dynamics to handle pressure control. At the end of the MD simulation, the final structure was employed for molecular docking calculations.

3.2. Library preparation and virtual screening

The small molecule library was prepared by collecting over 7 million compounds from commercially available Ambinter catalogue-2014 (www.ambinter.com) and publicly available ZINC (http://zinc.docking.org) [16], PubChem [4] and ChEMBL (www.ebi.ac.uk/chembl) [3] databases. Virtual screening was performed as described previously [2]. Before virtual screening, the library was subjected to the Filter-it software (version 1.0.2) to eliminate the compounds that violate Lipinski's rule of five [24] such as molecular weight, the number of hydrogen bond acceptors and donors, and logP value. Analysis by Filter-it software revealed 4 million compounds (4M library), which satisfy Lipinski's rule of five. For molecular docking calculations, docking software AutoDock Vina (version 1.1.2) which was available for public access was employed (http://vina.scripps.edu/) [37]. This version of AutoDock Vina predicts the optimal conformations of the receptor-ligand complex and report binding affinity scores by assuming a structure model with a rigid receptor (protein) and a flexible ligand.

3.3. Generation of molecular descriptors and feature extraction

Molecular descriptor values of the compounds in the small molecule library were generated by PaDEL-Descriptor [43] software. The software calculated the molecular descriptors developed by the Chemistry Development Kit (CDK) and provided 1875 descriptors (1444 1D, 2D descriptors and 431 3D descriptors) in 12 categories of fingerprints like atom type electrotopological state descriptors, Crippen's logP and MR, extended topochemical atom descriptors, McGowan volume, molecular linear free energy relation descriptors, ring counts and count of chemical substructures identified by Laggner. The selection of the best molecular descriptors for the training set, which defines low- and high-BFE classes (Fig. 2), was carried out for the MILP-HB. High number of repetitions in some of the descriptor values that possess the same value for at least 90% of the compounds were eliminated. Since the number of descriptors was much higher than the number of molecules, there was high multi-collinearity between the columns of the data set. We further reduced the descriptor set by using UFS [40] which chooses the maximal linearly independent columns with a minimal amount of multiple correlations by discarding the descriptors with standard deviation less than the predefined minimum standard deviation.

A PLSR model was constructed to select most informative descriptor set for the prediction of BFEs. PLSR was implemented by using the MINITAB statistical software (version 17) and the constructed model has an R^2 value of 0.95. In the PLSR model BFEs constitute the dependent variables (responses) and descriptor values are the independent variables. The 15 molecular descriptors (SC-5, SP-6, SHBd, minHaaCH, maxwHBa, FMF, MDEN-23, ndssC, nHBint2, ATSc2, C1SP3, minsCH3, nAtomP, BCUTp-1l, VC-5) with the highest regression coefficient were added to the '15 most significant descriptor' set, and the first 6 were used as input for the initial classification by MILP-HB.

3.4. MILP-HB classification and significance analysis

We used the MILP-HB method to classify the molecules. Each class is represented by several hyperboxes that enclose the molecules of the class and the hyperboxes of different classes do not overlap. The power of the method is due to its ability to construct multiple hyperboxes for each class that better fit the structure of the data. Basically, the existence of hyperboxes is represented by binary variables, yb_l takes the value 1 if box l exists and it is 0 otherwise. The boundaries of box l is indicated by continuous decision variables X_{lmn}, which is the bound n for box l on attribute (descriptor) m. The binary variable yp_{ik} represents the misclassification

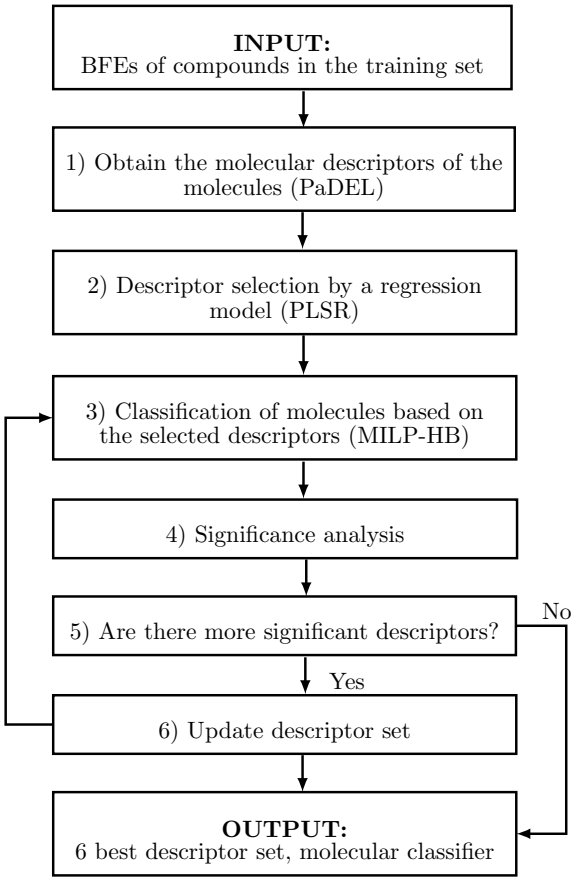


FIGURE 2. Schematic representation of the classification approach.

of sample i to class k. Sample data (set of molecules) is represented by the parameters, a_{im} which denotes the value of descriptor m of molecule i and D_{ik} which is the class k of molecule i. The mixed-integer programming model is constructed with these decision variables and supplementary variables along with relevant constraints minimizes the objective function in equation (3.1):

$$z = \sum_i \sum_k yp_{ik} + \sum_l yb_l. \tag{3.1}$$

Here, the purpose is to enclose the sample data by a set of hyperboxes for each class using the minimum number of hyperboxes and misclassified data. See [38] for the fundamental model. The efficiency of the method was improved in [8] by determination of problematic instances, *i.e.* instances covered by hyperboxes of multiple classes, finding seeds for problematic instances of each class, and elimination of intersections between hyperboxes. We used the approach in [8] and the mixed-integer programming model was built with the 6 most significant descriptors we obtained from PLSR and solved by the Cplex 12.3 solver [15] using Gamside [35]. The same approach has been successfully implemented in classifying drug molecules considering their IC50 values on target proteins [1, 12]. Additionally, inhibitors of human cytochrome P450 enzymes have been classified with respect to BFE and pIC_{50} using common 6 molecular descriptors [7].

The step following the initial MILP-HB classification is the significance analysis. It is possible to improve the initial classification accuracy by replacing the weakest descriptors in the 6 most significant descriptor set

used in MILP-HB with the most informative descriptors of the 15 most significant descriptor set. The impact of each descriptor is evaluated by comparing the within class variance of descriptor values with the variance of the whole set of molecules. Equation (3.2) given below exhibits the F distribution.

$$\frac{S_{ij}^2/\sigma_i^2}{S_{ik}^2/\sigma_i^2} = S_{ij}^2/S_{ik}^2 = f_{vn}. \tag{3.2}$$

F distribution is used for hypothesis testing where S_{ij}^2 is the variance of the set of values of descriptor i for the whole set of molecules and S_{ik}^2 is the variance of the set of values of descriptor i for molecule class k. The null hypothesis H_0: $S_{ij}^2 = S_{ik}^2$, claims that variance of molecule class k has the same variance with the whole molecule set for descriptor i. We expect in class variance to be smaller than the variance of the whole molecule set for a descriptor to be significant. Then, alternative hypothesis H_a: $S_{ij}^2 > S_{ik}^2$ claims that molecule class k has smaller variance than the whole set. For a strong descriptor we expect to have low p values both for low- and high-BFE classes.

3.5. Current classification tools in use

We compared the accuracy of the MILP-HB method with the classifiers in the WEKA data mining package [11]. Among the classifiers of WEKA that we considered, OneR (One Rule) is a very simple but accurate method. It builds a one-level decision tree, learns a rule from each attribute, and selects the rule having the smallest error rate as the one rule [41]. Another method called logit boost (additive logistic regression) is for boosting any of the classifiers that manage weighted data. It uses the logistic regression system for the learning process [12]. SMO (sequential minimal optimization) is a method to train a support vector classifier using polynomial kernels by breaking a large quadratic programming (QP) optimization problem into smaller QP optimization problems [33]. A Bayesian network is a directed acyclic graph, with nodes representing the variables and with probabilities attached to the edges. A Bayesian network is specified by the training run and is then used to perform inference maximizing the likelihood [14]. A naive Bayes classifier is based on an application of Bayes' theorem and the naive assumption that the variables are independent of each other. The classifier is trained and a parameter estimation of the probability model, *i.e.*, the independent features probability model, is performed using the maximum likelihood method. A threshold-based classifier puts an upper limit on a probability output through a distribution classifier such that the misclassification error is minimized [41]. A decision stump classifies the data according to a threshold value obtained by maximizing a likelihood function and usually utilizes a boosting algorithm [34]. K-star30 and LWR (locally weighted regression) are instance-based classifiers. In contrast to neural networks or decision trees, which use training to build global representations of their target functions, instance-based learning constructs query-specific local models, fitting the training instances in a neighborhood around the query point. K-star is an improved version of the k-nearest-neighborhoods method, and LWR assigns instance-based weights to be used in linear regression [6]. IB1 is listed as a lazy classifier, in the sense that it stores the training instances and does not really do any work until the classification time. IB1 is an instance based learner. It finds the training instance closest in Euclidean distance to the given test instance. IBk is a k-nearest-neighbor classifier that uses the same idea [41].

4. Data and results

4.1. Molecular dynamics simulation

An energy minimization and MD simulation with NAMD software was performed for the refinement of atom coordinates of SIRT6 (pdb id: 3K35) as described previously to bring the 3D structure of the SIRT6 at physiological conditions [2, 5]. This procedure is required to have a before docking calculation in order to calculate binding energies at realistic conditions. The simulation showed a slight increment after minimization up to approximately 0.4 ns, and then no significant deviations for the rest of the simulation were observed, which demonstrated the convergence of RMSD and the stability of the given structure within 1 ns time period. The final structure at the end of the MD simulation was employed for molecular docking calculations.

TABLE 1. Significance test results for the 6 most significant descriptors.

Descriptor	Class	Sample var.	p value
FMF	all data	0.005	
	low-BFE class	0.003	0.015
	high-BFE class	0.004	0.077
maxwHBa	all data	0.274	
	low-BFE class	0.033	0.001
	high-BFE class	0.498	0.994
minHaaCH	all data	0.019	
	low-BFE class	0.002	0.001
	high-BFE class	0.033	0.989
SHBd	all data	0.115	
	low-BFE class	0.105	0.371
	high-BFE class	0.127	0.664
SP-6	all data	1.969	
	low-BFE class	1.358	0.077
	high-BFE class	0.806	0.001
SC-5	all data	0.038	
	low-BFE class	0.032	0.249
	high-BFE class	0.034	0.317

4.2. MILP-HB classification and significance analysis

In order to start the classification of the molecules based on the high- or low-BFE values, we prepared a training dataset. First, virtual screening was performed for randomly chosen 1000 compounds from 4M library to calculate BFE values. Secondly, 100 compounds (50 of which have BFE values above the threshold value and the other 50 have BFE values below the threshold value) were randomly chosen and used as the training set for MILP-HB classification. BFE of NAD+ (−8.3 kcal/mol), which is a substrate of SIRT6, was used as a threshold to classify molecules as low-BFE (above −8.3 kcal/mol) or high-BFE (below −8.3 kcal/mol) sets. Finally, the 15 molecular descriptors with the highest regression coefficient of the PLSR model were added to the '15 most significant descriptor' set, and the first 6 were used as input for the initial classification by MILP-HB. Initial classification by MILP-HB resulted in 83.55% accuracy for low- and high-BFE sets.

The following step of the training MILP-HB classification is the significance analysis. It is possible to improve the prediction accuracy by replacing the least informative descriptors in the 6 most significant descriptor set used in MILP-HB with the strongest descriptors of the 15 most significant descriptor set. The impact of each descriptor was evaluated by comparing the within class variance of descriptor values with the variance of the whole set of molecules.

From a strong descriptor, we expect to have low p values both for low- and high-BFE classes. In the significance analysis, we compared the p values of the 6 most significant descriptor set with the p values of the remaining 9 significant descriptor set and replaced the weak descriptors with the stronger ones. MILP-HB was implemented with the updated descriptor set. If there was any improvement in classification accuracy, we continued with the new set, otherwise we stopped the iteration. Our initial classification with 10-fold cross validation provided the accuracy of 83.55%. In order to check if we could find stronger descriptors to get higher accuracy, we calculated p values both for the best 6 and remaining 9 descriptor sets as shown in Tables 1 and 2, respectively.

The p values of the descriptors SHBd and SC-5 among the 6 most significant descriptors are higher than 0.2 for both low- and high-BFE classes. Especially for the high-BFE class, p values are as high as 0.664 and 0.317 for SHBd and SC-5 respectively. In order to increase the classification accuracy, we replaced them with the ones in the remaining 9 significant descriptors. In this set, descriptors BCUTp-1I, nAtomP and ndssC have p values close to or less than 0.1 for high-BFE class which means that they may have higher classification power for high-BFE class than the descriptors SHBd and SC-5. Since the p value of BCUTp-1I for low-BFE class is very high,

TABLE 2. Significance test results for the remaining 9 significant descriptors.

Descriptor	Class	Sample var.	p value
VC-5	all data	0.005	
	low-BFE class	0.005	0.521
	high-BFE class	0.005	0.439
BCUTp-1l	all data	0.072	
	low-BFE class	0.103	0.932
	high-BFE class	0.044	0.025
nAtomP	all data	32.075	
	low-BFE class	32.634	0.539
	high-BFE class	23.330	0.107
minsCH3	all data	0.005	
	low-BFE class	0.005	0.381
	high-BFE class	0.005	0.327
C1SP3	all data	4.155	
	low-BFE class	3.844	0.389
	high-BFE class	4.108	0.493
nHBint2	all data	1.291	
	low-BFE class	1.111	0.286
	high-BFE class	1.478	0.720
ATSc2	all data	0.007	
	low-BFE class	0.006	0.293
	high-BFE class	0.008	0.690
ndssC	all data	1.376	
	low-BFE class	1.610	0.747
	high-BFE class	1.020	0.121
MDEN-23	all data	0.396	
	low-BFE class	0.453	0.714
	high-BFE class	0.347	0.307

TABLE 3. Selected 6 PaDEL descriptors for MILP-HB classification.

Descriptor	Description	Descriptor Java Class
SC-5	Simple cluster, order 5	ChiCluster
SP-6	Simple path, order 6	PaDELChiPath
SHBd	Sum of E-States for strong H bond donors	Electrotop.StateAtomType
minHaaCH	Min. atom-type H E-State	Electrotop.StateAtomType
maxwHBa	Max. E-States for weak H bond acceptors	Electrotop.StateAtomType
FMF	Complexity of a molecule	FMF

we discarded it and formed a new descriptor set as an input for MILP-HB by replacing SHBd and SC-5 with nAtomP and ndssC. The new descriptor set provided lower classification accuracy than the initial descriptor set. Hence, the classification algorithm terminates. We concluded that the 6 most significant descriptors that we selected by using PLSR provide the highest classification accuracy with the MILP-HB method. See Table 3 for a description of the selected PaDEL molecular descriptors.

4.3. Comparison of accuracy with current classification tools in use

We used the 6-best descriptor set and conducted classification analysis 10 times with the formerly selected 100-molecule set. Table 4 summarizes the average accuracies and standard deviations of both WEKA classifiers

TABLE 4. Comparison of MILP-HB with WEKA classifiers by 10-fold cross validation.

Classification Method	Average Accuracy (%)	Std. Dev.
MILP-HB	83.55	1.20
BayesNet	82.10	1.20
SMO	81.70	0.67
Logit Boost	79.90	1.29
NBTree	79.30	1.77
Decision Stump	78.30	2.06
Threshold	77.70	3.97
JRip	76.80	1.62
OneR	76.30	3.09
NaiveBayes	75.40	1.35
LWL	75.40	2.37
K-star	74.70	1.57
IBk	72.80	0.92

and the MILP-HB method. The MILP-HB method provided the highest accuracy for classifying the molecules as having high- and low-BFE values. Among the WEKA classifiers, BayesNet and SMO provided the highest classification accuracies that were higher than 80%. The MILP-HB method is the most reliable of all, having the same standard deviation with Bayesnet classifier but providing a higher accuracy level. This result indicates that MILP-HB is the most accurate classifier compared to classification methods available in WEKA that can be combined with virtual screening for the structure-based drug design.

4.4. Virtual screening and candidate inhibitors of sirt6

The docking software AutoDock Vina (version 1.1.2) was employed for molecular docking calculations. This version of AutoDock Vina predicts the optimal conformations of the receptor-ligand complex and reports binding affinity scores by assuming a structure model with a rigid receptor (protein) and a flexible ligand.

NAD+ is used as a cofactor by SIRT6 to transfer an acetyl group from its substrate proteins to the ADP-ribose moiety of NAD+; this cleaves the coenzyme and releases nicotinamide and O-acetyl-ADP-ribose [13]. Therefore, NAD+ binding pocket (active site) was selected as the target region for virtual screening. Then, the library of 2.4 million compounds were screened based on the BFEs. Top 10 hit compounds were determined and reported as the candidate inhibitors of SIRT6. BFEs of the selected candidates were in the range of -11.2 to -11.9 kcal/mol which were quite low compared with BFEs of SIRT6 true substrate NAD+ (-8.3 kcal/mol). Last, top 10 hits were further analyzed according to important characteristics such as docking positions, hydrogen bonding interactions with the residues at the NAD+ binding site and close proximity, van der Waals interactions, exposed surface area, and pocket occupancy. Their interactions with the residues at the NAD+ binding site are listed in Table 5 (see Tab. A.1 in the Appendix for molecular structure and formula of the compounds).

Virtual screening and detailed docking analysis revealed that critical residues which interact with NAD+ are Asp61, Phe62, Arg63, Trp69, Trp186, Ser214, Gln216 and Ile217. The SIRT6-NAD+ complex exhibits several polar interactions through residues Asp61, Arg63, Trp186, Gln216 and Ile217 which stabilizes the ligand-protein interaction (Fig. 3). Detailed analysis showed that the nicotinamide ring of NAD+ makes hydrogen bonds with Arg63 and Gln216. Additionally, compounds A1, A4, A9 and A10 were predicted to form a number of hydrogen bonds with different residues in the NAD+ binding active site, as observed in the SIRT6-NAD+ complex. Namely, such interactions were predicted to occur with Arg63 and Gln216. Moreover, the adenine part of NAD+ was predicted to form a pi-pi stacking interaction with Trp186 which also occurs with compounds A1, A2 and A6. Additionally, compound A2 exhibits another pi-pi stacking with Phe62. Interestingly, compounds A3, A7 and A8 do not make any polar interactions with SIRT6 but exhibit van der Walls interactions with residues Phe62, Trp69 and Trp186.

TABLE 5. BFEs and interacting residues of the top 10 hit compounds.

ID	BFE (kcal/mol)	Interacting Residues
NAD+	–8.30	Asp61, Phe62, Arg63, Trp69, Trp186, Ser214, Gln216, Ile217
A1	–11.60	Lys13, Asp61, Phe62, Arg63, Trp69, Hsd131, Trp186, Gln216
A2	–11.30	Arg63, Phe62, Trp69, Hsd131, Trp186, Ser214, Gln216
A3	–11.40	Phe62, Arg63, Trp69, Leu184, Asp185, Trp186, Thr213, Gln216
A4	–11.90	Asp12, Phe62, Arg63, Trp69, Val113, Trp186, Ser214, Gln216
A5	–11.30	Asp12, Asp61, Phe62, Arg63, Trp69, Val113, Trp186, Gln216
A6	–11.30	Lys13, Asp61, Phe62, Arg63, Trp69, Val113, Trp186, Gln216
A7	–11.40	Asp12, Phe62, Arg63, Trp69, Gln111, Trp186
A8	–11.60	Asp12, Phe62, Arg63, Trp69, Met155, Ile183, Trp186, Gln216
A9	–11.30	Asp12, Phe62, Arg63, Trp69, Val113, Leu184, Trp186, Gln216
A10	–11.80	Lys13, Phe62, Arg63, Trp69, Val113, Trp186, Ser214, Gln216

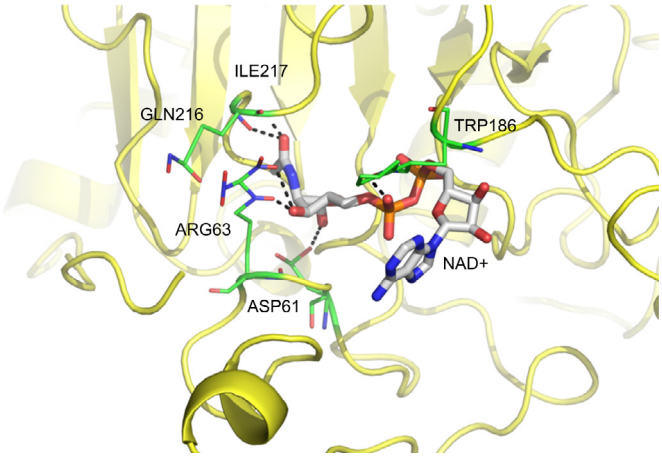


FIGURE 3. Interactions between NAD+ and SIRT6: NAD+ is shown as grey sticks, amino acid residue belonging to the SIRT6 are shown as green sticks (Asp61, Arg63, Trp186, Gln216 and Ile217). Hydrogen bonds are represented as black dotted lines.

5. CONCLUSION

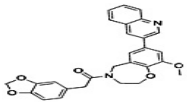
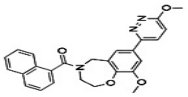
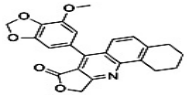
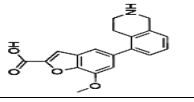
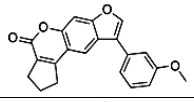
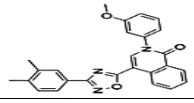
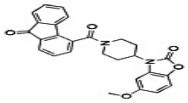
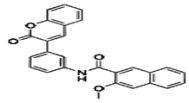
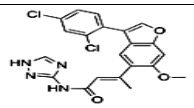
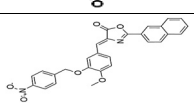
Virtual screening has become one of the major components in the structure-based drug discovery approach within the last few years. When the 3D structure of the protein is known, structure-based screening is preferred. However, virtual screening of chemical libraries with millions of compounds requires a lot of computing and data analysis time. Our motivation behind this study was to filter the initial small molecule libraries which contain compounds that may be deemed unsuitable for screening against a particular target protein by using a classification algorithm, MILP-HB, in combination with partial least squares regression. The effectiveness of the approach was illustrated on a target protein SIRT6. The availability of the crystal structure of SIRT6 makes structure-based drug design a viable approach to discover novel molecules that could inhibit the activity of the SIRT6 which plays an important role in metabolic disorders such as diabetes, in aging-related degenerative processes, and in cancer.

We constructed a PLSR model to select the 15 most informative descriptors among 1875 descriptors of PaDEL and by applying significance analysis we verified that the best 6 descriptor set provides the highest accuracy. We compared the average accuracies of MILP-HB and WEKA classifiers to justify the classification power of MILP-HB. The results indicated that the proposed approach outperformed other approaches reported in the

literature with 83.55% accuracy using the common 6 molecular descriptors (SC-5, SP-6, SHBd, minHaaCH, maxwHBa, FMF). As a final step, we constructed a MILP-HB classifier based on the whole set of 100 molecules so as to classify the molecules among the set of 4 million small molecules library to have high and low-BFE values and discard the high-BFE class of molecules for further consideration. 1.6 million molecules were classified as high-BFE and constitute 40% of the whole set. This further elimination of molecules by MILP-HB has provided us to reduce the computational time for virtual drug screening that requires a lot of time for the docking of millions of compounds. The remaining 2.4 million compounds in the low-BFE class were docked to SIRT6. Due to lack of any known inhibitors of SIRT6 in the literature, top 10 hit compounds were determined and reported as the candidate inhibitors of SIRT6. BFEs of the selected candidates were in the range of −11.2 to −11.9 kcal/mol. We concluded that the values of descriptors SC-5, SP-6, SHBd, minHaaCH, maxwHBa, FMF of an upcoming molecule suffice to predict its BFE class as high or low for SIRT6.

Appendix A

Table A.1. Molecular structure and formula of the compounds.

Compound ID	Formula	Structure
A1	C28H24N2O5	
A2	C26H23N3O4	
A3	C23H19NO5	
A4	C19H17NO4	
A5	C21H16O4	
A6	C26H21N3O3	
A7	C27H22N2O5	
A8	C27H19NO4	
A9	C21H16Cl2N4O3	
A10	C28H20N2O6	

References

[1] P. Armutlu, M.E. Ozdemir, F. Uney-Yuksektepe, I.H. Kavakli and M. Turkay, Classification of drug molecules considering their ic50 values using mixed-integer linear programming based hyper-boxes method. *BMC Bioinform.* **9** (2008) 411.

[2] P. Armutlu, M.E. Ozdemir, S. Ozdas, İ.H. Kavakli and M. Turkay, Discovery of novel cyp17 inhibitors for the treatment of prostate cancer with structure-based drug design. *Lett. Drug Design Discov.* **6** (2009) 337–344.

[3] A.P. Bento, *et al.*, The chembl bioactivity database: an update. *Nucleic Acids Research* **42** (2014) D1083–D1090.

[4] E.E. Bolton, Y. Wang, P.A. Thiessen and S.H. Bryant, Pubchem: integrated platform of small molecules and biological activities. *Ann. Rep. Comput. Chem.* **4** (2008) 217–241.

[5] B. Cakir, O. Dagliyan, E. Dagyildiz, I. Baris, I.H. Kavakli, S. Kizilel and M. Turkay, Structure based discovery of small molecules to regulate the activity of human insulin degrading enzyme. *PloS One* **7** (2012) e31787.

[6] J.G. Cleary and L.E. Trigg, K*: An instance-based learner using an entropic distance measure. In vol. 5 of *Proc. of the 12th International Conference on Machine learning* (1995) 108–114.

[7] O. Dagliyan, I.H. Kavakli and M. Turkay, Classification of cytochrome p450 inhibitors with respect to binding free energy and pic50 using common molecular descriptors. *J. Chem. Inf. Model.* **49** (2009) 2403–2411.

[8] O. Dagliyan, F. Uney-Yuksektepe, I.H. Kavakli and M. Turkay, Optimization based tumor classification from microarray gene expression data. *PloS One* **6** (2011) e14579.

[9] J.-P. Etchegaray, L. Zhong and R. Mostoslavsky, The histone deacetylase sirt6: at the crossroads between epigenetics, metabolism and disease. *Curr. Topics Med. Chem.* **13** (2013) 2991–3000.

[10] T. Finkel, C.-X. Deng and R. Mostoslavsky, Recent progress in the biology and physiology of sirtuins. *Nature* **460** (2009) 587–591.

[11] E. Frank, M. Hall, L. Trigg, G. Holmes and I.H. Witten, Data mining in bioinformatics using weka. *Bioinform.* **20** (2004) 2479–2481.

[12] J. Friedman, *et al.*, Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *Ann. Statist.* **28** (2000) 337–407.

[13] R.A. Frye, Phylogenetic classification of prokaryotic and eukaryotic sir2-like proteins. *Biochem. Biophys. Res. Commun.* **273** (2000) 793–798.

[14] D. Heckerman, A tutorial on learning with Bayesian networks. Springer (1998).

[15] IBM ILOG, Cplex user's manual 12.2 (2010).

[16] J.J. Irwin and B.K. Shoichet, Zinc-a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **45** (2005) 177–182.

[17] W.L. Jorgensen, The many roles of computation in drug discovery. *Science* **303** (2004) 1813–1818.

[18] P. Kahraman and M. Turkay, Classification of 1, 4-dihydropyridine calcium channel antagonists using the hyperbox approach. *Ind. Eng. Chem. Res.* **46** (2007) 4921–4929.

[19] A. Kaidi, B.T. Weinert, C. Choudhar, and S.P. Jackson, Human sirt6 promotes dna end resection through ctip deacetylation. *Science* **329** (2010) 1348–1353.

[20] Y. Kanfi, *et al.*, Regulation of sirt6 protein levels by nutrient availability. *FEBS Lett.* **582** (2008) 543–548.

[21] T.L.A. Kawahara, *et al.*, Sirt6 links histone h3 lysine 9 deacetylation to nf-κb-dependent gene expression and organismal life span. *Cell* **136** (2009) 62–74.

[22] D.B. Kitchen, H. Decornez, J.R. Furr and J. Bajorath, Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.* **3** (2004) 935–949.

[23] H. Kubinyi, Similarity and dissimilarity: a medicinal chemist's view. *Perspect. Drug Discov. Design* **9** (1998) 225–252.

[24] C.A. Lipinski, F. Lombardo, B.W. Dominy and P.J. Feeney, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **64** (2012) 4–17.

[25] G. Liszt, E. Ford, M. Kurtev and L. Guarente, Mouse sir2 homolog sirt6 is a nuclear adp-ribosyltransferase. *J. Biol. Chem.* **280** (2005) 21313–21320.

[26] A.D. MacKerell, *et al.*, All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **102** (1998) 3586–3616.

[27] Z. Mao, C. Hine, X. Tian, M. Van Meter, M. Au, A. Vaidya, A. Seluanov and V. Gorbunova, Sirt6 promotes dna repair under stress by activating parp1. *Science* **332** (2011) 1443–1446.

[28] E. Michishita, J.Y. Park, J.M. Burneskis, J.C. Barrett and I. Horikawa, Evolutionarily conserved and nonconserved cellular localizations and functions of human sirt proteins. *Mol. Biol. Cell* **16** (2005) 4623–4635.

[29] T. Mitchell and G.A. Showell, Design strategies for building drug-like chemical libraries. *Curr. Opin. Drug Discov. Devel.* **4** (2001) 314–318.

[30] R. Mostoslavsky, *et al.*, Genomic instability and aging-like phenotype in the absence of mammalian sirt6. *Cell* **124** (2006) 315–329.

[31] P.W. Pan, J.L. Feldman, M.K. Devries, A. Dong, A.M. Edwards and J.M. Denu, Structure and biochemical functions of sirt6. *J. Biol. Chem.* **286** (2011) 14575–14587.

[32] J.C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R.D. Skeel, L. Kale and K. Schulten, Scalable molecular dynamics with namd. *J. Comput. Chem.* **26** (2005) 1781–1802.

[33] J. Platt, Fast training of support vector machines using sequential minimal optimization. In vol. 3 of *Advances in Kernel Methods-Support Vector Learn.* (1999).

[34] Y. Qu, B.-L. Adam, Y. Yasui, M.D. Ward, L.H. Cazares, P.F. Schellhammer, Z. Feng, O.J. Semmes and G.L. Wright, Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients. *Clin. Chem.* **48** (2002) 1835–1843.

[35] R.E. Rosenthal, Gams – a user's guide (2015).

[36] M. Szachniuk, M.C. De Cola, G. Felicia and J. Blazewicz, The orderly colored longest path problem – a survey of applications and new algorithms. *RAIRO: OR* **48** (2014) 25–51.

[37] O. Trott and A.J. Olson, Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **31** (2010) 455–461.

[38] F. Uney and M. Turkay, A mixed-integer programming approach to multi-class data classification problem. *Eur. J. Oper. Res.* **173** (2006) 910–920.

[39] H.O. Villar and M.R. Hansen, Design of chemical libraries for screening. *Expert Opinion on Drug Discovery* **4** (2009) 1215–1220.

[40] D.C. Whitley, M.G. Ford and D.J. Livingstone, Unsupervised forward selection: a method for eliminating redundant variables. *J. Chem. Inf. Comput. Sci.* **40** (2000) 1160–1168.

[41] I.H. Witten and E. Frank, Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann (2005).

[42] S. Wold, M. Sjöström and L. Eriksson, Pls-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems* **58** (2001) 109–130.

[43] C.W. Yap, Padel-descriptor: An open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* **32** (2011) 1466–1474.

[44] L. Zhong, *et al.*, The histone deacetylase sirt6 regulates glucose homeostasis *via* hif1α. *Cell* **140** (2010) 280–293.