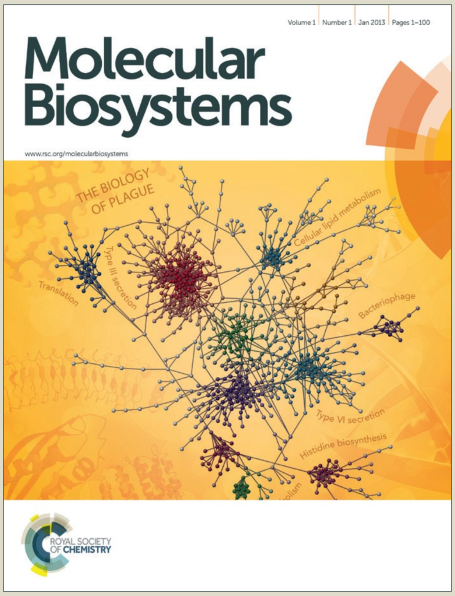
Volume 1 | Number 1 | Jan 2013 | Pages 1–100
Molecular Biosystems
www.rsc.org/molecularbiosystems
THE BIOLOGY OF PLAGUE
Cellular lipid metabolism
Type III secretion
Translation
Bacteriophage
Type VI secretion
Histidine biosynthesis
ROYAL SOCIETY OF CHEMISTRY

Molecular BioSystems

ARTICLE

An improved approach for predicting drug-target interaction: proteochemometrics to molecular docking

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

www.rsc.org/

Naeem Shaikh, Mahesh Sharma and Prabha Garg*

Proteochemometric (PCM) methods, which use descriptors of both the interacting species, i.e. drug and the target, are being successfully employed for the prediction of drug-target interactions (DTI). However, unavailability of non-interacting dataset and determining the applicability domain (AD) of model are main concern in PCM modeling. In the present study, traditional PCM modeling was improved by devising novel methodologies for reliable negative dataset generation and fingerprint based AD analysis. In addition, various types of descriptors and classifiers were evaluated for their performance. The Random Forest and Support Vector Machine models outperformed the other classifiers (accuracies >98% and >89% for 10-fold cross validation and external validation, respectively). The type of protein descriptors was having negligible effect on the developed models, encouraging the use of sequence-based descriptors over the structure-based descriptors. To establish the practical utility of built models, targets were predicted for approved anticancer drugs of natural origin. The molecular recognition interactions between the predicted drug-target pair were quantified with the help of reverse molecular docking approach. Majority of predicted targets are known for the anticancer therapy. These results thus correlate well with anticancer potential of the selected drugs. Interestingly, out of the predicted DTIs, thirty were found to be reported in ChEMBL database, further validating the adopted methodology. The outcomes of this study suggest that proposed approach, involving use of the improved PCM methodology and molecular docking, can be successfully employed to elucidate the intricate mode of action for drug molecules as well as repositioning them for new therapeutic applications.

Introduction

Polypharmacology is the phenomenon exhibited by drugs in which it binds to several macromolecular targets in complex biological system to exhibit the phenotypic effect.[1] Thus, identification of these interactions can help unfolding actual mechanism behind proposed therapeutic activity and side effects. It can also pave the way for drug repurposing. Therefore, determining complete interaction profile of pharmaceutical candidate against molecular targets is of prime importance for an efficient drug discovery.

Experimental ways to profile small molecules interactions with multiple targets are costly, time-consuming and practically impossible on such large scale. Hence, *in silico* drug-target interaction (DTI) prediction becomes a potential complement that provides useful information in an efficient way. The most straightforward approach is molecular docking which considers chemical interaction affinity to predict DTI. Though limited by speed, large-scale docking studies have been used to profile compounds against multiple targets.[2-4] Also, ligand-based approaches like Quantitative Structure–Activity Relationship (QSAR) have been widely used to predict activities of small molecules.[5] However, QSAR considers the interaction of single target with a group of compounds. Extension of QSAR for predicting activities of large group of compounds on large group of targets leads to the concept of proteochemometrics (PCMs). PCM modeling takes into account both drug and target information.[6] It can also take full advantage of multiple interaction data available and therefore can predict possible multiple interactions of a compound. PCM models can handle large amount of data, as they can be implemented on variety of machine-learning techniques.

PCM models have evidently been shown to outclass ligand-based models.[7, 8] PCM models were first reported for set of melanocortin receptors [9] and adrenergic G protein-coupled receptors.[10] Jamel Meslamani and Didier Rognan introduced three-dimensional (3D) binding site kernel along with standard chemical similarity kernel in support vector machine (SVM) classifier for better prediction of DTI.[11] The Random Forest (RF) and SVM based models were developed by integrating the chemical, genomic, and pharmacological information.[12] Current advancements and applications of PCM modeling has been extensively reviewed elsewhere.[13, 14]

Key limitation in PCM modeling is unavailability of large-scale negative dataset. Previous studies have used unlabeled

Department of Pharmacoinformatics, National Institute of Pharmaceutical Education and Research (NIPER), S. A. S. Nagar, Punjab 160062, India
**Corresponding Author: prabhagarg@niper.ac.in ; gargprabha@yahoo.com*

† Footnotes relating to the title and/or authors should appear here.

Electronic Supplementary Information (ESI) available: [details of any supplementary information available should be included here]. See DOI: 10.1039/x0xx00000x interactions as negative instances for model development. A rational methodology is needed for generation of negative instances to develop reliable predictive model. In addition, establishment of applicability domain (AD) of PCM models has not been considered in previous studies. AD analysis will be helpful in providing the confidence of prediction for the query molecule.

In this work, PCM models were developed for predicting DTI. Effect of machine-learning algorithms (classifiers) and descriptors on DTI prediction was also analyzed. A rational methodology for generation for reliable negative instance has been proposed. These models were extensively validated using external dataset extracted from various other bioactivity databases. Furthermore, a novel fingerprint based approach was devised to establish the AD for developed models. To prove the competence of models, protein targets for nature-derived anticancer drugs were predicted. Predicted interactions were further refined by molecular docking studies. Thus, in this work, a systematic approach for identifying drug-target interactions is formulated.

Material and methods

The structural database was selected to study the effect of sequence-based and structure-based descriptors on the model development and prediction. Co-crystallized structures were considered as interacting pairs where proteins were regarded as target and ligands as an interacting drug.

Training set

Interaction data was retrieved from the sc-PDB (v2011).[11] The sc-PDB is an annotated database of druggable binding sites gathered from the Protein Data Bank (PDB) containing 9877 entries (ligand-target interactions) of 3034 distinct proteins and 5339 different ligands. Ligands were inspected and prepared in Discovery Studio (v2.5.5).[15] Binding sites were annotated to UniProt accession number and complete protein sequences were retrieved from UniProtKB.[16] Pairwise distance matrix was prepared for these sc-PDB entries by using Euclidean distances, calculated on FuzCav binding site fingerprints (See Descriptor calculation). *AffinityPropagation* method of scikit-learn 0.15 [17] was used to cluster sc-PDB entries based on this pairwise distance matrix with 0.9 dumping factor. Clusters containing more than 10 entries (207 clusters) were retained.

Final dataset for training was composed of 3063 sc-PDB complexes comprising 1473 different targets and 2040 different ligands (See training_set.txt). Majority of training targets belong to *Homo sapiens*. Fig. 1 shows diversity across the clustered interactions used as training data for model development with respect to organism, enzyme class and resolution of crystalized structure.

External Validation set

External dataset for validation was collected from ChEMBL 17,[18] DrugBank 4.0,[19] IUPHAR [20] and Therapeutic Target Database (TTD).[21] All interactions for training targets were extracted from these databases. For retrieving data from ChEMBL, compounds with reported activities in terms of IC50, pIC50, Ki, pKi, Kd, pKd, EC50 and pEC50 were considered. Compounds with activity value equal or better than 1 μM against training protein targets with a confidence score of 8 or 9 were extracted. These rules were applied to ensure the selection of reliable bioactivity with high confidence target annotations. External dataset ligands were selected with molecular weight in-between of 140-800 Da. Tanimoto similarity with respective training ligands was kept less than 0.8. Finally, 91,566 non-redundant interaction covering 593 training targets and 72,504 new ligands were retrieved from these databases (Fig. S1).

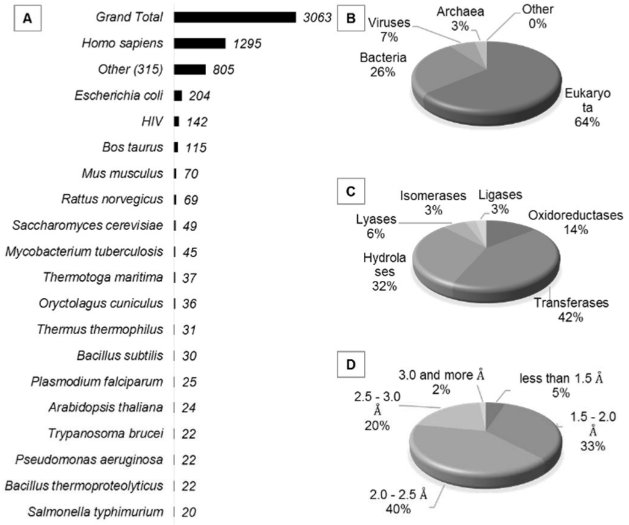

Fig. 1 Diversity analysis of the training set with respect to organism (A, B), enzyme classes (C) and resolution of crystal structure (D)

Negative sets

Target-ligand complexes from sc-PDB were considered as positive instances. Negative instances are also required to train supervised machine-learning model. Due to the unavailability of large-scale non-interacting pairs, a rationalized methodology was devised to generate reliable negative instances. Clean drug-like molecules at reference pH 7 were retrieved from ZINC database [22] for decoy set preparation. For each cluster of training targets, decoy ligands were selected to generate reliable negative instances as follows (Fig. 2):

1. All active ligands for each member of cluster target were retrieved from training and external test set.
2. Decoy ligands were selected from ZINC dataset if they are:
   a. Structurally dissimilar to any of the active ligands (Tanimoto similarity less than 0.2)
   b. Similar to active ligands with respect to five physicochemical descriptors *viz.* molecular weight, number of rotational bonds, total hydrogen bond donors, total hydrogen bond acceptors and the octanol–water partition coefficient (Euclidean distance less than 0.2) to remove artificial enrichment.
   c. Structurally dissimilar to other selected decoys (Tanimoto similarity less than 0.7) to ensure diversity in selection.
3. These decoys were then randomly paired with cluster targets to generate unlabeled instance for training and external test set.
4. For further refinement of training negative instances, weighted SVM classifier was trained with true pairs as positive instances (more weight) and unlabeled as negative instances.
5. Then this classifier was used to classify unlabeled instances.
6. Unlabeled instances, which were classified as positive *i.e.* false positives, were discarded.
7. New SVM model was then trained with positive and remaining unlabeled as negative instances.
8. Step (5), (6) and (7) were repeated until none of the unlabeled instance was predicted as positive.

Circular fingerprint (Morgan) was used to calculate structural similarity which was found to perform better in decoy generation methodology.[23] Physicochemically similar decoys were selected to make the dataset challenging and to remove artificial enrichment caused by strict topological dissimilarity cutoff. Decoy selection methodologies based on structural fingerprint and physiochemical properties have been used in previous studies.[23-25] The main problem associated with computationally generated negative set is the false decoys i.e. unknown positive/active instances in decoys. Activity of false decoy molecules can be attributed to the presence of active scaffolds or ligand binding warheads.[23]

Simply selecting topologically different molecules as negative instance might not be sufficient to remove false decoys, therefore a machine learning based approach has been used to further refine decoy set.[26] False decoys were iteratively removed by applying series of weighted SVM classifier (step 4-8). Higher weight (10:1) was applied on positive instances so that any decoy close to ligand can be removed. SVM models was trained with Morgan circular fingerprint as a descriptors, hence any decoys containing molecular fragment signature similar to that of active ligand will be predicted as active and removed from the decoy list. By virtue of this methodology, false decoys were identified and removed from decoy set to make reliable negative dataset. As a result, 11,796 and 182,505 reliable negative instances were generated for training set and external validation set, respectively (Fig. S1).

Morgan fingerprints (See Descriptor calculation), Tanimoto chemical similarity and physicochemical descriptors were calculated for each active and decoy ligands using RDKit 2015.03.1, a python toolkit (http://www.rdkit.org/). Euclidean distances between physicochemical descriptors of active and decoys were calculated in scikit-learn. SVM models were also developed in scikit-learn with RBF kernel, with two set of descriptors.

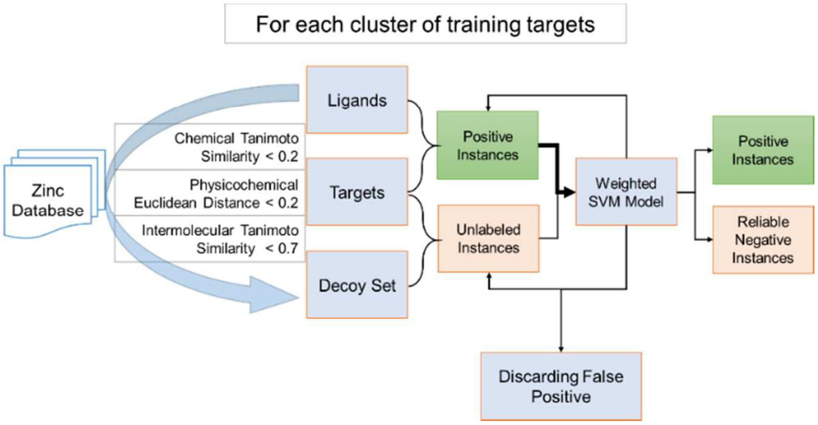
For each cluster of training targets
Zinc Database
Chemical Tanimoto Similarity < 0.2
Physicochemical Euclidean Distance < 0.2
Intermolecular Tanimoto Similarity < 0.7
Ligands
Targets
Decoy Set
Positive Instances
Unlabeled Instances
Weighted SVM Model
Positive Instances
Reliable Negative Instances
Discarding False Positive

Fig. 2 Methodology for negative instance generation

Descriptor calculation

Selection of descriptors was based on literature review. Mainly two set of protein descriptors were selected for study *viz.* sequence-based and structure-based descriptors to analyze their effect on the predictive ability. Total 1080 protein sequence descriptors were calculated using PROFEAT webserver.[27] Whereas, 3D structural descriptor of the binding site as implemented in FuzCav were used.[28] For each binding site, counts of all possible Pharmacophoric triplets within definite distance ranges between Cα carbon atoms were represented in the form of 4834 bit fingerprint. Only Morgan fingerprints were used as ligand descriptors as they were consistently found suitable in previous studies. These circular topological fingerprints are chemically interpretable and capture a large amount of information.[29] These are constructed by recording substructures around each non-hydrogen atom. The substructures are limited by a radius defined by a certain number of covalent bonds. Fingerprints were calculated in RDKit (http://www.rdkit.org/) with radius 2 and hashed to a 1024-bit string. All descriptors were scaled between zero and one with the help of *MinMaxScaler* method in scikit-learn 0.15.[17] Descriptors were selected based on *feature_importances* method as implemented in RF Classifier (Table S1).

Model development

Various types of classifier were tested against two set descriptors (discussed earlier) used in this study. SVM,[30] RF,[31] Naive Bayes [32] and *k*-Nearest Neighbor classifiers were used for model development as implemented in scikit-learn 0.15 for development of model.[17] Table 1 enlists types of models developed by their combinations of descriptors and classifiers and their abbreviation, which will be followed in this paper. For SVM classifier values of C and gamma were optimized by grid search as 10 and 0.001 respectively. Radial basis function was used as kernel for SVM classifier. The number of trees was kept 100 for RF classifier. Remaining parameters were kept as default. Number of neighbors were set to 1, 3 and 5 for *k*-Nearest Neighbors classifiers. Distance based weight was used so that closer neighbors of a query point will have a greater influence.

Table 1 List of models developed and their abbreviations

Classifier	Target Descriptors	
	Structure based	Sequence based
RF	3d-rf	2d-rf
SVM	3d-svm	2d-svm
Naive Bayes	3d-nb	2d-nb
k-Nearest Neighbors (k=1)	3d-knn1	2d-knn1
k-Nearest Neighbors (k=3)	3d-knn3	2d-knn3
k-Nearest Neighbors (k=5)	3d-knn5	2d-knn5

Model validation

Models were extensively validated by 10-fold cross validation and independent external validation. Following parameters were used for assessment of the prediction quality of models.

$$Recall\ (Sensitivity) = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Precision = \frac{TP + TN}{TP + TN + FP + FN}$$

$$f1 - score = 2 \times \frac{Recall \times Precission}{Recall + Precision}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(FP + TN)(TN + FN)}}$$

Where TP, FP, TN and FN represents true positive, false positive, true negative and false negative respectively. MCC denotes Matthews Correlation Coefficient. Recall and sensitivity are used to evaluate correct identification of positive and negative instances respectively. F1-score, the harmonic mean of recall and precision (positive predictive value) was used to measure the prediction accuracy. MCC measures the balanced prediction of binary classification models.[33, 34] Beside these parameters, Receiver Operating Characteristic Area Under the Curve (ROC_auc) score was also used as implemented in scikit-learn to illustrate the prediction performance.[17]

Molecular docking

To confirm and quantify the interaction of drug molecules to predicted target with respect to chemical interaction affinity, large-scale molecular docking studies were performed in Autodock Vina.[35] All 3063 binding sites and their respective ligands were prepared in pdbqt format. In contrast to conventional grid preparation method, the surface of ligand was considered instead of center of ligand. Grid was prepared by forming a 3D box around native ligand whose surfaces were 6.5 Å apart from edge of ligand in each direction. These ligands were redocked in respective binding sites for the purpose of comparison. Root Mean Square Deviation (RMSD) for co-crystal and redocked poses was calculated with python script. Average RMSD for best redock pose was found to be 1.48 Å, which validates the docking protocol. These results are comparatively better than previously reported high throughput docking studies.[36, 37] Furthermore redock RMSD for individual enzyme families and non-enzymes were also investigated (Table S2, Table S3). Average RMSD for each class was well within 1.7 Å. In addition, scores for co-crystal poses were also calculated by Autodock Vina for comparison. Python scripts were written to automate the large-scale docking and analysis of result.

Results and discussion

Model development and validation

Twelve models were developed by the combination of types of descriptors and classifiers. External validation set, comprises of 91,566 positive and 182,505 negative instances was employed for robust validation of models. All developed models were showing good accuracy i.e. more than 91% for 10-fold cross validation and more than 87% for external validation (Table 2). Best and nearly similar results were obtained by RF and SVM models with respect to ROC_auc score, accuracy and MCC (Fig. 3A1 and 3B1) irrespective of type of descriptors. However, Fig. 3A2 and 3B2 is suggesting that RF models are having slight advantage due to more balanced ratio of recall and specificity. Next best results were obtained by Naïve Bayes classifiers. Increasing the number of neighbor for *k*-Nearest Neighbor resulted in decrease in recall.

SVM and RF models were further analyzed for analogue biasness i.e. influence of prevalence of certain chemotypes in collected dataset.[38] New external dataset was developed by clustering initial external dataset on the basis of Murcko Scaffold.[39] Only one randomly selected representative of each cluster was used in new dataset. The new external dataset was composed of 41,161 positive and 108,858 negative instances. Interestingly, it was found that predictive performance for new external dataset was only insignificantly reduced as compared to initial external dataset prediction (Table S4). These results suggest that developed models are devoid of analogue biasness.

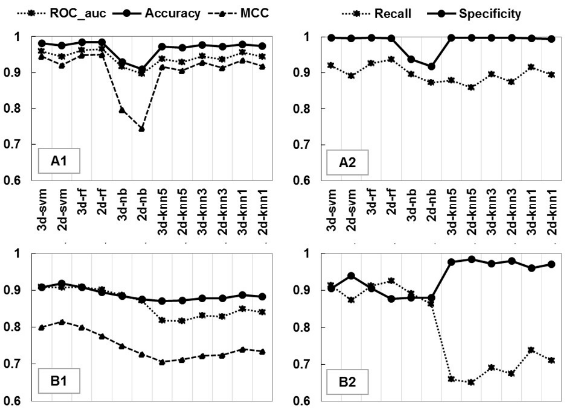

Fig. 3 Statistical parameters for 10-fold cross validation (A1, A2) and external validation (B1, B2)

Table 2 Statistical parameters for 10-fold cross validation and external validation

Dataset	Models	ROC_auc	Recall	Specificity	Precision	Accuracy	f1_score	MCC
10-fold Cross Validation	3d-svm	0.96	0.92	1.00	0.99	0.98	0.96	0.95
	2d-svm	0.94	0.89	1.00	0.99	0.98	0.94	0.92
	3d-rf	0.96	0.93	1.00	0.99	0.98	0.96	0.95
	2d-rf	0.97	0.94	1.00	0.98	0.98	0.96	0.95
	3d-nb	0.92	0.90	0.94	0.79	0.93	0.84	0.80
	2d-nb	0.90	0.87	0.92	0.74	0.91	0.80	0.75
	3d-knn5	0.94	0.88	1.00	0.99	0.97	0.93	0.92
	2d-knn5	0.93	0.86	1.00	0.99	0.97	0.92	0.91
	3d-knn3	0.95	0.90	1.00	0.99	0.98	0.94	0.93
	2d-knn3	0.94	0.87	1.00	0.99	0.97	0.93	0.91
	3d-knn1	0.96	0.92	1.00	0.98	0.98	0.95	0.94
	2d-knn1	0.94	0.89	0.99	0.98	0.97	0.93	0.92
External Validation	3d-svm	0.91	0.91	0.91	0.83	0.91	0.87	0.80
	2d-svm	0.91	0.87	0.94	0.88	0.92	0.88	0.82
	3d-rf	0.91	0.91	0.91	0.83	0.91	0.87	0.80
	2d-rf	0.90	0.93	0.88	0.79	0.89	0.85	0.78
	3d-nb	0.89	0.89	0.88	0.79	0.88	0.84	0.75
	2d-nb	0.87	0.86	0.88	0.78	0.88	0.82	0.73
	3d-knn5	0.82	0.66	0.98	0.93	0.87	0.77	0.71
	2d-knn5	0.82	0.65	0.98	0.95	0.87	0.77	0.71
	3d-knn3	0.83	0.69	0.97	0.93	0.88	0.79	0.72
	2d-knn3	0.83	0.68	0.98	0.94	0.88	0.79	0.73
	3d-knn1	0.85	0.74	0.96	0.90	0.89	0.81	0.74
	2d-knn1	0.84	0.71	0.97	0.92	0.88	0.80	0.74

Applicability domain analysis

Ideally, in order to have practical application, a prediction model should be accompanied with analysis of AD. As the methodology for AD analysis of fingerprint-based model is not well defined, a novel approach has been devised for determining applicability domain for the developed model. AD analysis was performed for better performing RF models. RF is an ensemble of decision trees, which uses the presence or absence of structural features in the training set compounds to predict the category of a test compound. Therefore, it can be inferred that RF model predicts the activity based on bit combinations present in the training set compounds. The absence of such bit combinations in the query molecule will not lead to reliable prediction by the developed model. Thus, it can be hypothesized that presence of training set bit-pairs in the query molecule can be used as a measure of AD analysis.

To test this hypothesis, external set was evaluated for the shared bit combinations. For this purpose, bit-pairs pool were generated by combinatorial pairing of important bit features for each training compound (important bit features were previously identified by RF; See Descriptor calculation). For each external set molecule, combinatorial bit-pairs were generated and number of shared bit pairs with the generated pool was identified. External set was then categorized with respect of number of common bit pairs (bin size = 25). Each category was subjected to predictions using the developed RF model and then recall was calculated. The Fig. 4 and Table S5 shows count of external set molecule and recall against number of common bit pairs. It can be clearly concluded that the molecules with lower number of common bit pairs are poorly predicted thus validating hypothesis.

The AD analysis indicates that number of common bit pairs can be efficiently used to estimate the confidence of prediction by the model. PCM models cover wide variety of chemical space, thus bound to have large applicability domain as can be seen in Fig. 4. However, a query molecule should contain at least a few of the bit-pairs used in training the models. The number of common pair is proportional to the recall. For a reliable prediction (>70% recall), at least 75 common bit-pairs are needed.

Target prediction for nature-derived anticancer drugs

To demonstrate the applicability of the models, 137 anticancer drugs derived from natural resources [40] were subjected to target prediction by the RF models. The reason behind selecting this dataset was the observation that the mode of action of both natural product based drugs and anticancer drugs is poorly understood and therefore have scope for identification of novel therapeutic targets. Drugs having more than 800 Da molecular weight were discarded, as it was the maximum limit for training set molecules. Drugs with less than 75 common bit-pairs were also removed to make the dataset within the AD of model prediction. Cutoff of prediction probability was set to be 0.65 for each RF model. Limit for average prediction probability (APP) was set to be 0.70. Top 25 predictions for each drug by each model were retrieved and redundancy was removed. These predicted targets were further filtered for therapeutic applications. Only those targets, which have been reported in TTD of OMIM database (http://www.omim.org/), were considered for further studies. These predictions were additionally validated by molecular docking studies in Autodock Vina (Fig. 5). These drugs were docked into the binding cavities of predicted targets. Native ligands of these targets were also docked into respective cavities for comparison. Only those predictions were considered which were having comparable docking score with the native ligand (*i.e.* docking score of predicted docking less than -8.0 kcal/mol or difference with native ligand docking score not more than 2 unit).

Final set of prediction was composed of 264 complexes for 54 drug molecules (See Case_Study_results.txt). Following cases were used to exemplify the usefulness of proposed methodology for target prediction. Details of docking poses can be found in supplementary information. (Fig. S2-S3)

Predicting known targets. ChEMBL database was searched for targets with smiles query for drug molecules. Total 35 predicted interactions were already reported in ChEMBL, out of which 30 were unseen by the model (See ChEMBL_match.txt). These results further validate the PCM models and target prediction methodology. For example, out of 11 predicted targets for Sorafenib, seven targets were found in ChEMBL to have binding affinity with Sorafenib. Remarkably, three of newly predicted targets were known to have anticancer activity (Fig. 6).

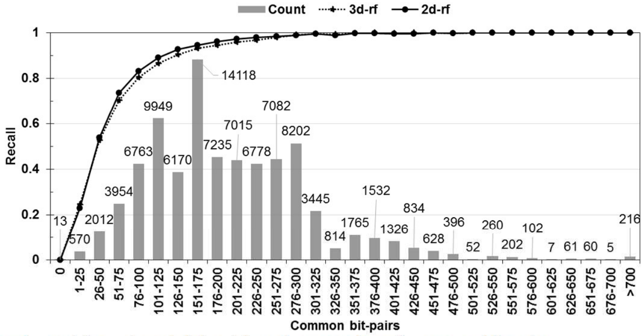

Fig. 4 Recall and count for external dataset based on number of common bit-pairs

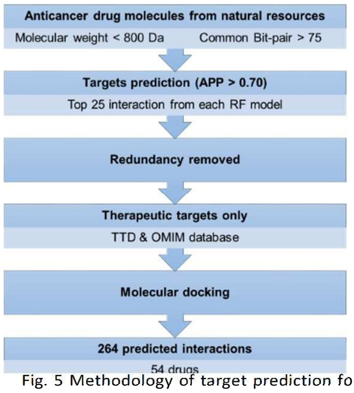
Anticancer drug molecules from natural resources
Molecular weight < 800 Da
Common Bit-pair > 75
Targets prediction (APP > 0.70)
Top 25 interaction from each RF model
Redundancy removed
Therapeutic targets only
TTD & OMIM database
Molecular docking
264 predicted interactions
54 drugs

Fig. 5 Methodology of target prediction for case study

Predictions supporting anticancer activity. Nearly half of the predicted interactions relate drugs to targets, which are involved in cancers. These newly predicted targets may help in understanding the intricate mechanism of action of anticancer drug molecules. For example, Table 3 shows the predicted targets for Nadrolone phenylpropionate.

All predicted targets for Nadrolone phenylpropionate belongs to *Homo sapiens* and have been reported to play an important role in cancer pathogenesis and therapeutics.[41-43] X-ray crystal structure of QR2 with resveratrol and inhibitory activity of steroidal pyrazolines against it advises the activity of compounds with steroidal scaffold against QR2.[44] All these findings suggest that these predicted targets may be involved in anticancer activity of Nadrolone phenylpropionate. Docking poses for these interactions can be found in supplement information (Fig. S2).

Predictions supporting other activities. Many of the predicted targets were involved in diseases other than cancer. These predictions provide opportunity to repurpose drug molecules for other therapeutic activity. For example, Table 4 shows predicted targets for Colchicine and Demecolcine. PDE4D, Cruzipain, Dipeptidyl peptidase-4 were predicted for both Colchicine and its derivative. PDE4B and PDE4D are involved in psoriasis and other skin related conditions. This prediction could be linked to anti-psoriatic activity of Colchicine. Fig. 7 represents docking poses for Demecolcine in PDE32 and PDE43. Aromatic ring with alkoxy groups of Demecolcine are greatly overlapping with that of native ligands in respective esterases. π-π stacking interaction with phenylalanine and hydrogen bonding with glutamine residue are also present in docked poses. In addition, Demecolcine is forming salt bridge with adjacent aspartate residue that was not present in case of native ligand.

Colchicine and Demecolcine were also predicted to act against malarial targets. Previously, Colchicine has been found to be active against *Plasmodium falciparum,* though not preferred as an anti-malarial therapy due to its cytotoxicity.[45] These predictions suggest that antimalarial activity of Colchicine and their derivatives can also be attributed to targets other than tubulin, which include Cruzipain and Enoyl reductase. Therefore, designing Colchicine derivatives selective for Cruzipain and Enoyl reductase could be beneficial for non-toxic antimalarial activity. Interaction with Glucosylceramidase can explain the activity of Colchicine against metabolic disorder. Similarly, interaction with Beta-secretase 1 and Dipeptidyl peptidase are pointing out towards its probable application for Alzheimer’s disease and diabetes mellitus.[46]

As an interesting observation, all anthracyclines have been predicted with good APP and docking score against Aldose Reductase 1B1 (Table 5, Fig. S3). AKR1B1 is the mediator of inflammatory signals induced by growth factors, cytokines, chemokines, carcinogens etc.[47] These interactions might be playing a crucial role in anticancer activity of anthracyclines. It also suggests other therapeutic application for anthracyclines associated with aldose reductase such as rheumatoid arthritis, diabetic nephropathy etc

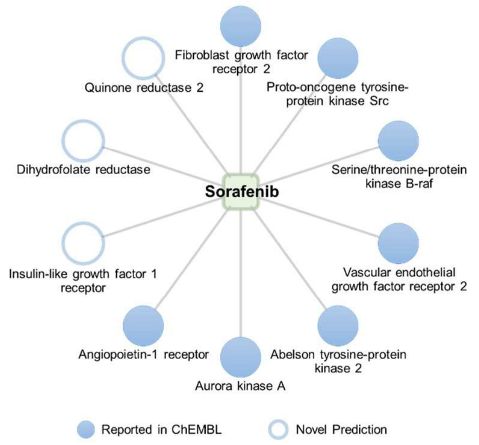
Fibroblast growth factor receptor 2
Quinone reductase 2
Proto-oncogene tyrosine-protein kinase Src
Dihydrofolate reductase
Sorafenib
Serine/threonine-protein kinase B-raf
Insulin-like growth factor 1 receptor
Vascular endothelial growth factor receptor 2
Angiopoietin-1 receptor
Abelson tyrosine-protein kinase 2
Aurora kinase A
● Reported in ChEMBL
○ Novel Prediction

Fig. 6 Predicted anticancer targets for Sorafenib

Table 3 Predicted targets for Nadrolone phenylpropionate

Predicted target	APP	Dock Score
Hepatocyte growth factor receptor	0.92	-9.3
Kinesin-like protein KIF11	0.92	-9.0
Focal adhesion kinase 1	0.89	-8.6
Quinone reductase 2	0.89	-11.4
Abelson tyrosine-protein kinase 2	0.87	-9.5

Dock score in kcal/mol

Table 4 Predicted targets for Colchicine and Demecolcine

Generic Name	APP	Dock score	Protein names	Organism
Colchicine	0.83	-7.0	cAMP-specific 3',5'-cyclic phosphodiesterase 4D (PDE43)	*Homo sapiens*
	0.77	-5.2	Cruzipain	*Trypanosoma cruzi*
	0.79	-6.4	Dipeptidyl peptidase 4	*Homo sapiens*
	0.78	-7.9	Glucosylceramidase	*Homo sapiens*
Demecolcine	0.79	-7.4	Beta-secretase 1	*Homo sapiens*
	0.76	-7.8	cAMP-specific 3',5'-cyclic phosphodiesterase 4B (PDE32)	*Homo sapiens*
	0.79	-8.0	cAMP-specific 3',5'-cyclic phosphodiesterase 4D (PDE43)	*Homo sapiens*
	0.74	-5.1	Cruzipain	*Trypanosoma cruzi*
	0.78	-6.3	Dipeptidyl peptidase 4	*Homo sapiens*
	0.76	-6.8	Enoyl-ACP reductase	*Plasmodium falciparum*
	0.72	-6.1	Quinone reductase 2	*Homo sapiens*

Dock score in kcal/mol

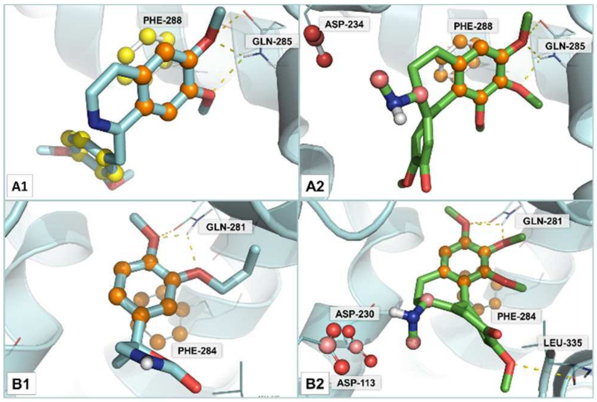

Fig. 7 Docking poses for Demecolcine (A1) and native ligands (A2) in PDE43; Demecolcine (B1) and native ligands (B2) in PDE32

Table 5 Anthracyclines against with aldose reductase 1B1

Generic Name	APP	Drug score
Amrubicin	0.92	-9.3
Daunorubicin	0.83	-9.6
Doxorubicin	0.83	-8.5
Epirubicin	0.83	-9.9
Idarubicin	0.85	-9.8

Dock score in kcal/mol

Conclusions

In this work, an improved PCM based approach is reported in order to predict the interaction profile of given chemical moiety against various therapeutic targets. A foremost problem in PCM modeling is the unavailability of non-interacting dataset, which is required as negative instance for training machine-learning model. Previous studies have employed unknown interactions as negative instances, which is not a correct representation of non-interacting dataset. Therefore, to develop reliable model, a novel approach has been devised in this work for selecting non-interacting pairs. Another significant development of this work is the establishment of AD based on Morgan fingerprints, which are 6known to capture large amount of chemical information. The presence of at least 75 common fingerprint bit pairs in the query molecule and training set increases the confidence on the model predictions. The RF-based models marginally outperformed the SVM-based models with respect to performance and speed. The use of structure-based descriptors was not having any significant advantage over sequence-based descriptors. These results encourage the use of sequence-based descriptors for model development, as it will lead to increase the coverage of dataset. Developed models deliver predictions for over 1473 protein targets. Though it does not cover the entire human proteome, this is a large number of protein targets.

To unveil the practical application of the built models, target prediction was carried out for a dataset of approved anticancer drugs derived from natural resources. Large-scale molecular docking studies were performed to further quantify and delineate the actual molecular recognition interaction between drugs and predicted targets. From the case studies, it can be concluded that many predicted targets were relevant to the known anticancer and other activity of drug molecules. These predicted targets can provide means to explore actual mechanism of anticancer activity of drug molecules. Many of the known DTI that were unseen by the models were also correctly identified, further validating the prediction capability of developed models. Novel identified targets can be explored to extend the activity profile of drug molecules to new therapeutic classes, thus enabling repurposing. In the wider perspective, this methodology indeed provides a useful way to identify plausible targets for small molecules. Addressing such polypharmacological profile of pharmaceutical candidate will influence the drug discovery pipeline positively.

Acknowledgements

This work is financially supported by Department of Electronics and Information Technology, GOI, New Delhi, India, (DIT/R&D/Bio/15(3)/2011).

References

1. J. Mestres, E. Gregori-Puigjané, S. Valverde and R. V. Solé, Mol. BioSyst., 2009, 5, 1051-1057.
2. H. Patel, X. Lucas, I. Bendik, S. Günther and I. Merfort, ChemMedChem, 2015, 10, 1209-1217.
3. H. Li, Z. Gao, L. Kang, H. Zhang, K. Yang, K. Yu, X. Luo, W. Zhu, K. Chen and J. Shen, Nucleic Acids Res., 2006, 34, W219-W224.
4. V. Belekar, A. Shah and P. Garg, Mol. Divers., 2013, 17, 97-110.
5. D. Sprous, R. Palmer, J. Swanson and M. Lawless, Curr. Top. Med. Chem., 2010, 10, 619-637.
6. J. E. S. Wikberg, O. Spjuth, M. Eklund and M. Lapins, in Computational Approaches in Cheminformatics and Bioinformatics, John Wiley & Sons, Inc., 2011, DOI: 10.1002/9781118131411.ch3, pp. 57-92.
7. N. Weill and D. Rognan, J. Chem. Inf. Model., 2009, 49, 1049-1062.
8. H. Geppert, J. Humrich, D. Stumpfe, T. Gärtner and J. Bajorath, J. Chem. Inf. Model., 2009, 49, 767-779.
9. P. Prusis, R. Muceniece, P. Andersson, C. Post, T. Lundstedt and J. E. Wikberg, BBA-Protein Struct. M., 2001, 1544, 350-357.
10. M. Lapinsh, P. Prusis, A. Gutcaits, T. Lundstedt and J. E. Wikberg, BBA-Gen. Subjects, 2001, 1525, 180-190.
11. J. Meslamani and D. Rognan, J. Chem. Inf. Model., 2011, 51, 1593-1603.
12. H. Yu, J. Chen, X. Xu, Y. Li, H. Zhao, Y. Fang, X. Li, W. Zhou, W. Wang and Y. Wang, PLoS ONE, 2012, 7, e37608.
13. I. Cortés-Ciriano, Q. U. Ain, V. Subramanian, E. B. Lenselink, O. Méndez-Lucio, A. P. IJzerman, G. Wohlfahrt, P. Prusis, T. E. Malliavin and G. J. van Westen, MedChemComm, 2015, 6, 24-50.
14. G. J. van Westen, J. K. Wegner, A. P. IJzerman, H. W. van Vlijmen and A. Bender, MedChemComm, 2011, 2, 16-30.
15. D. Studio, Accelrys Inc.: San Diego, CA, USA, 2009.
16. C. The UniProt, Nucleic Acids Res., 2015, 43, D204-D212.
17. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss and V. Dubourg, J. Mach. Learn. Res., 2011, 12, 2825-2830.
18. A. P. Bento, A. Gaulton, A. Hersey, L. J. Bellis, J. Chambers, M. Davies, F. A. Krüger, Y. Light, L. Mak and S. McGlinchey, Nucleic Acids Res., 2014, 42, D1083-D1090.
19. V. Law, C. Knox, Y. Djoumbou, T. Jewison, A. C. Guo, Y. Liu, A. Maciejewski, D. Arndt, M. Wilson and V. Neveu, Nucleic Acids Res., 2014, 42, D1091-D1097.
20. J. L. Sharman, H. E. Benson, A. J. Pawson, V. Lukito, C. P. Mpamhanga, V. Bombail, A. P. Davenport, J. A. Peters, M. Spedding and A. J. Harmar, Nucleic Acids Res., 2012, 41, D1083-D1088.
21. F. Zhu, Z. Shi, C. Qin, L. Tao, X. Liu, F. Xu, L. Zhang, Y. Song, X. Liu and J. Zhang, Nucleic Acids Res., 2011, 40, D1128-D1136.
22. J. J. Irwin and B. K. Shoichet, J. Chem. Inf. Model., 2005, 45, 177-182.
23. M. M. Mysinger, M. Carchia, J. J. Irwin and B. K. Shoichet, J. Med. Chem., 2012, 55, 6582-6594.
24. N. Huang, B. K. Shoichet and J. J. Irwin, J. Med. Chem., 2006, 49, 6789-6801.
25. I. Wallach and R. Lilien, J. Chem. Inf. Model., 2011, 51, 196-202.
26. B. Zhang and W. Zuo, Journal of Computers, 2009, 4, 94-101.
27. H. Rao, F. Zhu, G. Yang, Z. Li and Y. Chen, Nucleic Acids Res., 2011, 39, W385-W390.
28. N. Weill and D. Rognan, J. Chem. Inf. Model., 2010, 50, 123-135.
29. D. Rogers and M. Hahn, J. Chem. Inf. Model., 2010, 50, 742-754.
30. V. Vapnik, The nature of statistical learning theory, springer, 2000.
31. L. Breiman, Machine learning, 2001, 45, 5-32.
32. H. Zhang, AA, 2004, 1, 3.
33. P. Garg, R. Dhakne and V. Belekar, Mol. Divers., 19, 163-172.
34. N. S. H. Narayana Moorthy and V. Poongavanam, RSC Advances, 2015, 5, 14663-14669.
35. O. Trott and A. J. Olson, J. Comput. Chem., 2010, 31, 455-461.
36. M. X. LaBute, X. Zhang, J. Lenderman, B. J. Bennion, S. E. Wong and F. C. Lightstone, 2014.
37. X. Zhang, S. E. Wong and F. C. Lightstone, J. Comput. Chem., 2013, 34, 915-927.
38. A. C. Good and T. I. Oprea, J. Comput-Aided. Mol. Des., 2008, 22, 169-178.
39. G. W. Bemis and M. A. Murcko, J. Med. Chem., 1996, 39, 2887-2893.
40. F. Zhu, C. Qin, L. Tao, X. Liu, Z. Shi, X. Ma, J. Jia, Y. Tan, C. Cui and J. Lin, Proc. Natl. Acad. Sci. U. S. A., 2011, 108, 12943-12948.
41. T. Y. Seiwert, T. N. Beck and R. Salgia, in Molecular Determinants of Head and Neck Cancer, Springer, 2014, pp. 91-111.
42. D. Bongero, L. Paoluzzi, E. Marchi, K. M. Zullo, R. Neisa, Y. Mao, R. Escandon, K. Wood and O. A. O'Connor, Leuk. Lymphoma, 2014, 1-25.
43. E. K. Greuber, P. Smith-Pearson, J. Wang and A. M. Pendergast, Nat. Rev. Cancer, 2013, 13, 559-571.
44. M. M. Abdalla, M. A. Al-Omar, M. A. Bhat, A.-G. E. Amr and A. M. Al-Mohizea, Int. J. Biol. Macromol., 2012, 50, 1127-1132.
45. K. Na-Bangchang and J. Karbwang, Fundam. Clin. Pharmacol., 2009, 23, 387-409.
46. A. Bhat, S. M. Naguwa, G. S. Cheema and M. E. Gershwin, Ann. N. Y. Acad. Sci., 2009, 1173, 766-773.
47. R. Tammali, S. K. Srivastava and K. V. Ramana, Curr. Cancer Drug Targets, 2011, 11, 560.