J Mol Model (2009) 15:591–596
DOI 10.1007/s00894-008-0349-1

ORIGINAL PAPER

kNNsim: k-Nearest neighbors similarity with genetic algorithm features optimization enhances the prediction of activity classes for small molecules

Dariusz Plewczynski

Received: 12 April 2008 / Accepted: 2 July 2008 / Published online: 29 July 2008

Abstract Protein targets specificity classification is an important step in computational drug development and design efforts. The enhanced classification models of small chemical molecules enable the rapid scanning of large compounds databases. Here, we present the k-nearest neighbors with genetic algorithm feature optimization approach for selection of small molecule protein inhibitors. The method is trained on selected, diverse activity classes of the MDL drug data report (MDDR) with ligands described using simple atom pairs two dimensional chemical descriptors. The accuracy of inhibitors identification is presented in confusion tables with calculated recall and precision values. The precision for selected types of targets exceeded 70%, and the recall reaches 40%. As a consequence, the method can be easily applied to large commercial compounds collections in a drug development campaign in order to significantly reduce the number of ligands for further costly experimental validation.

Keywords Chemical descriptors · Compound identification · Machine-learning methods · MDL drug data report · Protein target specificity · Substrate specificity · Virtual high-throughput screening

Introduction

The similarity of drug targets is typically measured using sequence or structural information. The similarity of their inhibitors can also be measured by comparing their structure, or some chemical characteristics. Here, we consider chemoinformatics approach that uses ligand similarity in order to find new, potent small chemical molecules that can inhibit a given protein target. This similarity between molecules can be mapped in a form of a network. The biological activity of those molecules can be used to divide the network into activity clusters, therefore enabling the actual prediction of inhibition for novel small molecules. Derived in that way the ligands networks are very robust to changes in chemoinformatics metrics, and such they can be used for reliable prediction of pharmacology [1]. Those chemoinformatics networks are stable to the method used to calculate the ligand-set similarities and to the chemical representation of the ligands. The ligand based networks were found to be small-world and broad-scale [1]. In the same way proteins can be grouped into functional groups based on their inhibitors similarity calculated using ligands' topology. Relating receptors by ligand chemistry show unexpected relationships that may be then verified by experiments using the ligands themselves [2].

This general principle that maps the chemical structure similarity onto similar biological activity is applied in very different applications. For example the preclinical safety pharmacology (PSP) attempts to anticipate adverse drug reactions (ADRs) during early phases of drug discovery by testing compounds on their contribution to known clinical ADRs (clinical trials, animal experiments, and molecular studies). In the work of Benter et al. the detailed analysis of pharmacology data was performed, and the prediction of adverse drug reactions and off-target effects from chemical structure only [3]. They explored PSP chemical space and its relevance for the prediction of adverse drug reactions.

D. Plewczynski (✉)
Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw,
Pawinskiego 5a,
02-106 Warsaw, Poland
e-mail: D.Plewczynski@icm.edu.pl

They constructed *in silico* Bayesian models for 70 PSP-related targets, which are able to detect 93% of the ligands binding to those targets at the classification rate of about 94% [3]. Then they employed the World Drug Index (WDI) and built the models for adverse drug reactions based only on normalized side-effect annotations in the WDI, without any analysis of the underlying functional knowledge. On average 90% of the adverse drug reactions observed with known, clinically used compounds were detected, an overall correct classification rate of 92% [3]. Combining PSP and ADR models enable authors to propose new hypotheses linking targets and adverse effects even without precise knowledge about the structure of a protein target.

Similarly, the high-throughput screening (HTS) performs activity testing of millions of compounds for chosen protein target. Therefore it can identify an initial set of lead molecules with high probability of activity. Selected compounds are further prioritized by optimization of various molecular characteristics. Unfortunately this approach is very expensive, and cannot be applied on whole genomes scale. On the other hand recent advances in structural genomics provide an ever growing number of protein structures. The reduction of the number of compounds to be tested experimentally is needed in order to allow for wider system biology studies. This *in silico* approach is called virtual high-throughput screening (vHTS), and historically in most cases simple similarity searches were performed. Such computational approach uses only information about the structure of the inhibitor, without taking into account the protein structure, or underlying molecular mechanism of the interaction between protein and ligand. This is in principle similar to the preclinical safety pharmacology techniques as described in [3]. On the other hand, when the structure of a protein target is available, it can be used for guiding the similarity approaches even when the initial set of known inhibitors is very small, or empty [4, 5].

Recently machine learning algorithms were used for target specific compounds classification [5, 6]. Several machine learning techniques were trained to distinguish between kinases' inhibitors and non-active small chemical molecules [7]. They reviewed support vector machines, the k nearest neighbor classification with GA-optimized feature selection, the neural networks and recursive partitioning. Trend vector analysis in combination with topological descriptors, has proved useful in drug discovery for ranking large collections of chemical compounds in accordance with their biological activity classes [8, 9]. We present here the application of supervised machine learning algorithm, namely k-nearest neighbor with genetic algorithm feature optimization. The evaluation of our method is done for five divergent activity classes of the highest medicinal relevance, which already have been investigated in several drug discovery programs or computational approaches. The selected activity classes are taken from the commercially available MDL drug data report [10]. Each ligand from the database can be described using various types of chemical descriptors. The simplest descriptors, like 2D structure of ligand, allow for a partial estimate of ligand activity for a given protein target [5]. In most of machine learning approaches, as reviewed by Plewczynski et al. [11], the use of parameters describing the compound's topology gave satisfactory results. Bender and Glen [12] using the number of atoms per element were able to outperform virtual-affinity-based fingerprints and unity fingerprints in some activity classes. Therefore here, we utilize also very simple atom pairs descriptors to describe small chemical molecules.

In our previous work we performed the wide evaluation of different machine learning methods [11]. The support vector machines, random forest, artificial neural networks, k-nearest-neighbor, naive Bayesian classification, and decision tree were used there to identify the active compounds for a selected protein target. Previously we reported differences in the overall performance of different methods depending on the biological target and activity class. Different methods can have different applications; some provide particularly high enrichment, others are strong in retrieving the maximum number of actives [11].

In the present study we focus on k-nearest-neighbor method for identification of inhibitors of proteins with genetic algorithm optimized features trained on selected five activity classes of MDL drug data report [10]. kNNsim tool perform rapid screening of very large databases of small chemical molecules and selecting new ligands for known activity classes. We provide here an in-depth description and detailed results for k-nearest neighbor method. The presented methodology is able to perform virtual high-throughput screening, when the size of the database is the main obstacle. The comparison of our method with other machine learning approaches is outside of the scope of this manuscript. kNNsim tool allows for faster scanning of large ligands databases in comparison to other recently published methods, such as combination of SVM with naïve Bayesian trained on Ghose-Crippen parameters and others [5, 6, 7, 12, 13].

Method

First, we have selected five divergent activity classes from the MDL drug data report [10]. All compounds are clinically tested or already launched on the market. For each protein target known inhibitors were used for training k-nearest neighbor method with genetic algorithm feature

selection and optimization. Additional tests were performed on biologically tested compounds from the same database. The cyclooxygenase-2 activity class contains 112 inhibitors that were divided randomly into two subsets: training (75 inhibitors with randomly selected from other activity classes 2106 inactive ones) and testing (respectively 37 and 8346). The dihydrofolate activity class has 28 known inhibitors (divided into 17 for training and 11 for testing), and randomly selected 10529 inactive ones (dividing them to two groups: 2149 and 8380). In the case of reverse transcriptase we used 114 inhibitors (79 and 35) and 10450 inactive compounds (2130 and 8320), and for thrombin we have found 112 inhibitors (77 and 35) and 10459 inactive ones (2036 and 8423). For the antiestrogen inhibition class we have collected 34 inhibitors (22 and 12) and 11580 inactive compounds (2528 for training and 9052 for testing). In addition we have selected biologically tested molecules for inhibition of cyclooxygenase-2 (792 molecules), dihydrofolate (154), thrombin (1066), reverse transcriptase (597) and antiestrogen (256) protein targets.

Cheminformatics methods operate under the assumption that similar chemicals have similar biological activity. This principle bridges chemical and biological space, and is the key to drug discovery and development [14]. One could predict a ligand's biological activity given only its chemical structure by similarity searching in libraries of compounds with known activities. Yet the optimal selection of a similarity metric in chemical space depend on a particular protein target. The work of Nettles et al. compares both 2D and 3D chemical descriptors as tools for predicting the biological targets of ligand probes, on the basis of their similarity to reference molecules. The 2D methods in general outperform the 3D methods (88% vs 67% success) in protein target prediction [14]. These findings support idea to test similar chemical descriptors in the context of inhibition prediction using compounds from MDL MDDR database of known drugs.

The simplest two dimensional topological descriptors, i.e., the regular atom pair AP descriptors [15], were used in this study. This type of chemical descriptors have been proven to be successfully in classification of compounds for various activity classes. They are easy to use and interpret. Descriptors were calculated for all ligands using the MIX tool [16], which counts for each atom pair the number of covalent bonds that join them. Atom pairs represent each molecule as a binary vector with '1' for all present types of atom pairs, and '0' for those that are absent. We have tested also other chemical descriptors (such as TT regular topological torsion, DP pairs using sq types, DT torsions using SQ types, DRUGBITS substructures and ROF6 set of descriptors [16]), and detailed comparison of results is presented in Table 1.

Table 1 kNNGA precision and recall values for selected five MDDR activity classes trained using launched and preclinical inhibitors

Target	Precision	Recall
COX2	0.66	0.37
DH	0.1	0.25
TH	0.79	0.44
RT	0.42	0.26
AE	0.53	0.17

The results were obtained using a larger set of chemical descriptors, in comparison to Table 1, when only atom pairs were used for describe inhibitors.

For activity class prediction we used here the supervised k-nearest neighbors method (kNN) [17] that subdivides a set of input cases (characterized by the vectors of descriptors) into different classes. The kNN predicts a classification for test cases based on the majority voting of its k nearest neighbors in the feature space. In our implementation k=5 is used. We used the Euclidian metric for calculating distances and the same set of descriptors as for other methods. The most discriminatory descriptors are calculated using a genetic algorithm with four generations and 40 chromosomes [18].

The variable selection k-nearest neighbours (kNN) method is a typical nonlinear methodology for building quantitative structure-activity relationship (QSAR). It is based on calculating correlation between chemical descriptors of compounds and their biological activities. The activity models are trained by finding a subspace of the original descriptor space where activity of each compound in the data set is most accurately predicted as the averaged activity of its k nearest neighbors in this subspace [19]. In a different approach variable selection were to find active analogues, i.e., similar compounds that may display similar profiles of pharmacological activities. The activity of each compound is predicted as the average activity of K most chemically similar compounds from the data set [20]. The chemical structures are characterized by multiple topological descriptors such as molecular connectivity indices or atom pairs. The chemical similarity is evaluated by Euclidean distances between compounds in multidimensional descriptor space, and the optimal subset of descriptors is selected using simulated annealing as a stochastic optimization algorithm.

Performance

The performance of our supervised machine learning classifier is described here using accuracy E, precision P and recall R values, together with confusion tables. The

error estimates are calculated using the leave-one-out procedure using the following equations:

$$E = \frac{fp + fn}{tp + fp + tn + fn} * 100\%,$$

$$R = \frac{tp}{tp + fn} * 100\%,$$

$$P = \frac{tp}{tp + fp} * 100\%,$$

where *tp* is the number of true positives, *fp* is the number of false positives, *tn* is the number of true negatives and *fn* is the number of false negatives. The classification error E provides an overall error measure, whereas recall R measures the percentage of correct predictions (the probability of correct prediction), and precision P gives the percentage of observed positives that are correctly predicted (the measure of the reliability of positive instances prediction).

On Fig. 1 we present recall and precision values for the largest activity classes from MDDR database. We present the results of training kNNsim on the pre-clinical and launched compounds from MDDR, and comparing them to the results of training on the whole set of positives including biologically testing compounds. Similarly we present also the testing results on pre-clinical and launched ligands and for all positives. The precision and recall for all types of activity classes is over 70% for all protein targets on training sets, whereas on testing datasets it is slightly lower.

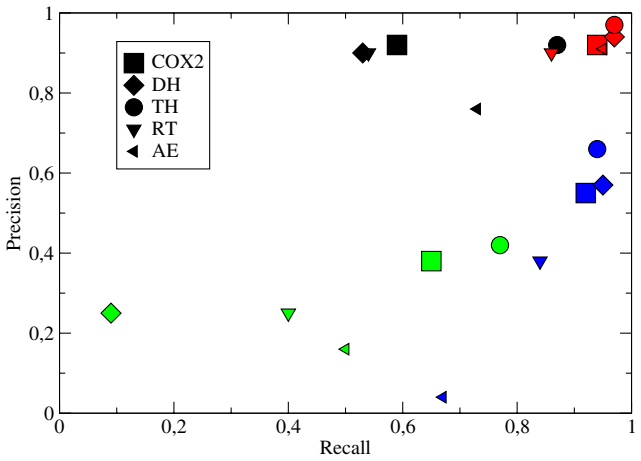

Fig. 1 The recall and precision values for five diverge MDDR activity classes. The black symbols mark the training results on the pre-clinical and launched compounds from MDDR. The red color is used for results of training on the whole set of positives including biologically testing compounds. The light green denotes testing results on pre-clinical and launched ligands. The blue is testing results for all positives. The precision and recall for all types of targets are over 70% for selected targets on training sets, whereas on testing datasets is lower

In Table 2 we present detailed benchmark results using confusion tables. The precision for selected types of targets exceeded 70%, and the recall reaches 40%. In the first table kNNsim precision and recall values are calculated for the five selected activity classes of MDDR database, namely cyclooxygenase-2 (COX2), dihydrofolate (DH), thrombin (TH), reverse transcriptase (RT) and antiestrogen (AE). The training was performed on two thirds of launched and preclinical inhibitors of those sets of compounds. For each activity class the accuracy, the classification error, the precision and recall values were calculated on the rest, i.e., one third of preclinical and launched inhibitors.

Those results are in close agreement with other comparative studies [7, 11, 21, 22]. The kNNsim method outperforms other types of QSAR methods in terms of its speed and relative precision. The high average precision (~70%), i.e., a number of false positives in the data set predicted to be active, suggest that kNN method can be used in low-throughput screening experiments in which only a few compounds can be validated. On the other hand the if one wants to retrieve actives as completely as possible, kNN can be a less convenient choice, as its average recall value (~40%) is below the values of other QSAR methods. The detailed results shows that k-nearest neighbors similarity with genetic algorithm features optimization is able to predict recently published ligands for a given protein target on the basis of initial leads, which is of crucial importance in medicinal research.

Conclusions

We present here a rapid software tool for selection of inhibitors from the large collection of compounds using some prior knowledge. The supervised machine learning approach, namely k-nearest neighbor with genetic algorithm feature optimization, is trained here on known members of several activity classes from MDDR database. Then the classification models are used for prediction of new molecules active for selected protein drug targets. Similarly, the method is able to enrich a set of known activites, when some prior knowledge is available [11]. When no initial information is available, one can use the results of docking experiment performed on a randomly selected subset of the whole database (for example 10%), in order to extend its results for the large commercial

Table 2 kNNGA precision and recall values for selected five MDDR activity classes trained using launched and preclinical inhibitors

All protein targets			
Target	Precision	Recall	
COX2	0.65	0.38	
DH	0.09	0.25	
TH	0.77	0.42	
RT	0.40	0.25	
AE	0.5	0.16	
Cyclooxygenase-2:			
	Predicted	0	1
Observed			8320
0	0	8347	**8307**
1	1	37	13
All	8384		
Accuracy		99.37%	
Classification error	0.63%		
Recall		64.86%	
Precision		37.50%	
Dihydrofolate:			
	Predicted	0	1
Observed		8397	4
0	8390	**8387**	3
1	11	10	**1**
All	8401		
Accuracy		99.85%	
Classification error	0.15%		
Recall		9.09%	
Precision		25.00%	
Thrombin:			
	Predicted	0	1
Observed		8395	65
0	8425	**8387**	38
1	35	8	**27**
All	8460		
Accuracy		99.46%	
Classification error	0.54%		
Recall		77.14%	
Precision		41.54%	
Reverse transcriptase:			
	Predicted	0	1
Observed		8302	56
0	8323	**8281**	42
1	35	21	**14**
all	8358		
Accuracy		99.25%	
Classification error	0.75%		
Recall		40.00%	
Precision		25.00%	
Antiestrogen:			
	Predicted	0	1
Observed		9027	38
0	9053	**9021**	32
1	12	6	**6**
all	9065		
Accuracy		99.58%	
Classification error	0.42%		
Recall	50.00%		
Precision	15.79%		

The list of protein targets include: cyclooxygenase-2 (COX2), dihydrofolate (DH), thrombin (TH), reverse transcriptase (RT) and antiestrogen (AE). Next tables present classification performance for each protein target calculated on the set of preclinical and launched inhibitors from MDDR database. Columns represent observed in experiments class of a compound for each of targets (active/inactive) whereas rows represent the prediction results. Here we present results on testing datasets with one third available positives (preclinical or launched inhibitors for the selected target) and two thirds of negatives (randomly selected subset of preclinical or launched inhibitors knowing not to inhibit selected target). The last four lines for each table present the calculated accuracy of the classification, the classification error, and the precision and recall values on the testing datasets using launched and pre-clinical compounds.

compounds collection. We showed that in this way we are able to recover 50% of all known actives for selected activity classes [5]. The kNNsim algorithm performs fast and reliable prediction of activity classes for previously unclassified compounds. It performs the classification of small molecules using 2D topological descriptors with respect to their potential inhibition on selected target classes. The MIX tools chemical descriptors [16] are useful for the different types of classification. The selection of molecular descriptors should be done in accordance with the balance between general and detailed level of description.

In our previous publication [11] we have compared different QSAR methods. We were interested in estimating the performance of those methods on standardized benchmark dataset both in terms of inhibition classes and the set of used molecular descriptors. In total seven different QSAR methods were trained to classify a diverse set of protein targets of medicinal importance. We have pointed out the significant differences in the performance of QSAR methods independent of the biological target and compound class. We have concluded, that different methods can have different applications; some provide particularly high precision, others are strong in retrieving the maximum number of active molecules.

In the present work we focused our attention on kNN, i.e., k-nearest neighbors similarity with genetic algorithm features optimization. We have shown that it enhances the efficiency of prediction of activity classes for small molecules for a wide set of protein targets. The kNNsim supervised machine learning method can be used for the virtual high-throughput screening campaign. It can prioritize the typically very large number of hits from screening

experiments. It also identifies compounds from large compounds collections for further experimental validation. In the context of the pharmaceutical industry it allows for designing new compounds that are not present in the present screening collection to be synthesized, or bought from a commercial source. The kNNsim can be therefore used in high-throughput fashion for solving practical problems of virtual screening. It is hoped that the new generation of *in silico* predictive models for drug activity is able to help support early QSAR to accelerate drug discovery. In addition, models such as the kNN based can be used for compound profiling in all development stages. Due to its relative simplicity, high degree of automation, nonlinear nature, and computational efficiency, this method could be applied routinely to a large variety of experimental data sets.

Acknowledgements This work was supported by EC within BioSapiens (LHSG-CT-2003–503265) and SEPSDA (SP22-CT-2004–003831) 6FP projects and the Polish Ministry of Education and Science (PBZ-MNiI-2/1/2005 and MNII ordinary research grant to DP).

References

1. Hert J, Keiser MJ, Irwin JJ, Oprea TI, Shoichet BK (2008) J Chem Inf Model 48(4):755–765
2. Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK (2007) Nat Biotechnol 25(2):197–206
3. Bender A, Scheiber J, Glick M, Davies JW, Azzaoui K, Hamon J, Urban L, Whitebread S, Jenkins JL (2007) ChemMedChem 2 (6):861–873
4. Ji ZL, Wang Y, Yu L, Han LY, Zheng CJ, Chen YZ (2006) Toxicol Lett 164(2):104–112
5. Plewczynski D, von Grotthuss M, Spieser SA, Rychlewski L, Wyrwicz LS, Ginalski K, Koch U (2007) Comb Chem High Throughput Screen 10(3):189–196
6. Fang J, Dong Y, Lushington GH, Ye QZ, Georg GI (2006) J Biomol Screen 11(2):138–144
7. Briem H, Gunther J (2005) Chembiochem 6(3):558–566
8. Sheridan RP, Nachbar RB, Bush BL (1994) J Comput Aided Mol Des 8(3):323–340
9. Wilton D, Willett P, Lawson K, Mullier G (2003) J Chem Inf Comput Sci 43(2):469–474
10. MDL, MDL Drug Data Report (2004) Coverage: 1988-present; updated monthly. Focus: Drugs launched or under development, as referenced in the patent literature, conference proceedings, and other sources; descriptions of therapeutic action and biological activity; tracking of compounds through development phases. Size: 132726 molecules,129459 models. Updates add approximately 10,000 new compounds per year. 2004
11. Plewczynski D, Spieser SA, Koch U (2006) J Chem Inf Model 46 (3):1098–1106
12. Bender A, Glen RC (2005) J Chem Inf Model 45(5):1369–1375
13. Nidhi, Glick M, Davies JW, Jenkins JL (2006) J Chem Inf Model 46(3):1124–1133
14. Nettles JH, Jenkins JL, Bender A, Deng Z, Davies JW, Glick M (2006) J Med Chem 49(23):6802–6810
15. Sheridan RP (2000) J Chem Inf Comput Sci 40(6):1456–1469
16. Miller MD, Sheridan RP, Kearsley SK (1999) J Med Chem 42 (9):1505–1514
17. Kauffman GW, Jurs PC (2001) J Chem Inf Comput Sci 41 (6):1553–1560
18. Raymer ML, Sanschagrin PC, Punch WF, Venkataraman S, Goodman ED, Kuhn LA (1997) J Mol Biol 265(4):445–464
19. Itskowitz P, Tropsha A (2005) J Chem Inf Model 45(3):777–785
20. Zheng W, Tropsha A (2000) J Chem Inf Comput Sci 40(1):185–194
21. Burbidge R, Trotter M, Buxton B, Holden S (2001) Comput Chem 26(1):5–14
22. Byvatov E, Fechner U, Sadowski J, Schneider G (2003) J Chem Inf Comput Sci 43(6):1882–1889