Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 57 (2015) 418 – 427

Procedia
Computer Science

3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015)

Ligand-Based Virtual Screening using Random Walk Kernel and Empirical Filters

Preeja M P[a], Hemant Palivela[b], Dr. K P Soman[c], Dr. Prashant S Kharkar[d]

[a,b]Department of Information Technology, MPSTME, NMIMS, Mumbai-400056, India
[c]Professor and Head, Department of Computational Engineering and Networking, Amruta University, Coimbatore - 641112, India
[d]Associate Professor and Head, Department of Pharmaceutical Chemistry, SPTM, NMIMS, Mumbai-400056, India

Abstract

Drug discovery is a time-consuming and costly process. The data generated during various stages of the drug discovery is drastically increasing and it forces machine-learning scientist to implement more effective and fast methods for the utilization of data for reducing the cost and time. Molecular graphs are very expressive which allow faster implementation of the machine-learning algorithms. During the discovery phase, virtual or *in silico*screening plays a major role in optimizing the synthesis efforts and reducing the attrition rate of the new chemical entities (NCEs). In the present work, a combination of the virtual screening using walk kernel and empirical filters was tried.The model was applied to two classification problems to predict mutagenicity and toxicity on two publically-available datasets. The accuracies obtained were 67 % for the PTC dataset and 87% for the MUTAG dataset. The results obtained from the combined method were found to be more accurate with less computational cost.

Peer-review under responsibility of organizing committee of the 3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015)

Keywords : Drug discovery, Virtual screening, Graph kernel, ADMET, SVM

1. Introduction

Drug discovery is the process of finding a new chemical compound with desired chemical and biological properties[1]. Two types of screening are involved in the discovery phase – *in silico*/virtual screening (VS), followed by experimental high-throughput screening (HTS). In the HTS, large numbers of compounds from various sources – in-house compound databases, commercial vendors, combinatorial libraries, natural product libraries, etc. - are quickly assayed using automated platforms; but the cost is very high [2]. The VS can complement the HTS by reducing the number of compounds to be screened, with increased probability of success. The cost-factor for VS is substantially low.

Virtual screening methods are classified into ligand-based and structure-based methods. In ligand-based strategy, the existing information about the ligand molecules is used for virtual screening of the designed or

Peer-review under responsibility of organizing committee of the 3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015)
doi:10.1016/j.procs.2015.07.508

physical compound collections. For structure-based strategy, the information about both target and ligand is used for the 'design cycle'. Ligand-based VS can be further classified into empirical filters-based, similarity-based and quantitative-structure activity relationship (QSAR)-based methods.

In QSAR-based method, the accuracy of screening depends on the way of representation of the molecules (molecular descriptors) and the method for finding the similarity or dissimilarity. Machine-learning methods play a big role in the development of accurate QSAR models. Recent research shows that kernel methods such as support vector machine (SVM) give more accurate QSAR models. In case of SVM, the classifier, the decision is based on the accuracy of similarity measure which in turn, depends on the representation as well as the similarity calculation. One-diimensional (1D), 2D and 3D representations of molecules (with increasing accuracy and computational complexity) can be used for virtual screening [3]. The 2D representation such as molecular graphs is simple and expressive due to which it is well suited for the faster development of graph-matching algorithms.

In the present work, ligand-based VS was used. The models were developed and then the implementation of the model and the performance of the algorithm on several different benchmark datasets were evaluated.

2. Ligand -based Virtual Screening

In this work, two types of ligand-based VS screening were implemented. In the first phase, random walk kernel-based similarity prediction was used for the classification of compounds into two sets, active and inactive. This was based on the QSAR modeling. In the second phase, the accuracy of model was increased by removing non-drug like molecules from the set of active compounds classified in the first phase. This was done by using filters based on Lipinski's rule-of-five and a similar rule-of-three.

2.1. QSAR-based Virtual Screening

Machine-learning methods are routinely used to establish QSAR. SVM is widely used because it gives promising result for non-linear QSAR problems. Compared to a majority of molecular descriptors for QSAR, molecular graphs include most of the information without much complexity. In SVM, the classification is based on the similarity calculation between molecules. The complexity of molecular graph comparison is very high and graph kernels are used for finding the molecular graph similarity. In this work, bond information-based molecular similarity methods were used.

2.1.1 Kernel Computation for Edge-Weighted Molecular Graphs

In the first phase, random walk kernel for edge-weighted molecular graphs was used for establishing QSAR. Here, the random walk kernel introduced in [5] was used. In this method, molecular similarity calculation is based on the connectivity between the atoms.

$$k(G,G') = \frac{1}{|G||G'|}\sum_{k=0}^{n} \boldsymbol{\lambda}^{k}\, \boldsymbol{e}^{T} A_{x}^{k} \boldsymbol{e} \qquad (1)$$

Here A_x is the adjacency matrix of product graph and $\boldsymbol{\lambda}$ is the weightage for different length of walks. The adjacency matrix of product graph is calculated using Kronecker product $A_X = A \otimes A',\ where\ A\ and\ A'$ are the adjacency matrices of graphs $G\ and\ G'$, respectively. This kernel is modified using edge weight matrix W_x. The edge weight matrix was used for kernel computation in [6]. Edge weight matrix is found out using adjacency matrix and the information about the edge labels.

Modified random walk kernel for edge weighted molecular graph

$$k_{walk}(G,G') = \frac{1}{|G||G'|}\sum_{k=0}^{n} \boldsymbol{\lambda}^k\, \boldsymbol{e}^T W_x^k \boldsymbol{e} \tag{2}$$

Normalized walk kernel

$$k'_{walk}\,(G_1,G_2) = \frac{k_{walk}\,(G_1,G_2)}{\sqrt{k_{walk}(G_1,G_1)*k_{walk}(G_1,G_1)}} \tag{3}$$

The similarity between graphs $G\ and\ G'$ was calculated using the above equation.

In QSAR, the aim was to classify molecules into active and inactive sets. SVM-based binary classifier was used for classifying molecules into active and inactive and m. In SVM, the decision of classifier depends on the value of the decision function

$$f(x) \;= sign(w^T x - \gamma) \tag{4}$$

where k'_{walk} was calculated using equation 3. w represents the coefficient of classifier obtained by solving the maximum margin optimization problem in SVM.

2.1.2 Empirical Filters for Drug-like Molecules

In phase 1, screening was based on the random walk kernel defined in Eq. 3. In phase 2, the empirical filters were used for the virtual screening of non drug-like molecules from the active set of compounds obtained from phase 1. In this work, empirical filters based on Lipinski RO5 and RO3 were used in [7,8]. Due to these empirical filters, sensitivity and specificity of classification were increased. Type 1 error occurs when inactive molecule is classifying as active molecule and type 2 error occurs when active molecule is classifying as inactive molecule. More importance is given for type 1 than type 2 error for the obvious reasons.

- Molecular mass: 180 to 500 Dalton
- High lipophilicity (expressed as LogP less than 5)
- Less than 5 hydrogen bond donors(OHs and NHs)
- Less than 10 hydrogen bond acceptors (Ns and Os)
- Molar refractivity should be between 40-130 [8]
- Exclude compound containing elements like Arsenic (As), Cadmium (Cd), Lead (Pb), and Mercury (Hg)

Drug-like molecule should not violate more than one rule from Lipinski rules. The rule 6 is included for removing toxic compounds from the active set in [9,10,11].

3. Implementation

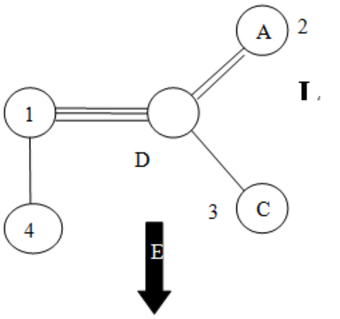
All the codes were written in Java Release 7, NetBeans IDE 7.3.1 and experiments were run on a 1.8 GHz Intel Core™ i3 processor with 4 GB of main memory running Ubuntu 13.04. After the computation of kernel matrix, the classification code was written in MATLAB Release 15. Classification was done with the help of LIBSVM version 2.88. In the second phase, the properties were calculated using E-Dragon and Interactive ALOGPS VCCLAB site [8,9]. Codes for empirical filters were also written in JAVA Release 7, NetBeans IDE 7.3.1. The practical suitability of these kernels were tested on two real-world datasets, PTC [13] and MUTAG [14]. These two datasets were repeatedly used for classification benchmarking [3, 4, 15, 16, 17, 18].

The random walk kernel between the molecules was generated from edge-weighted molecular graphs. The similarity calculation was based on the connectivity as well as the edge labels. In this work, edge-weighted molecular graphs were implemented by using type of bond as edge label. Walk kernel calculates similarity based on the linear features (e.g., sequence like C-C chain). Walks were extracted using adjacency matrix. Adjacency matrix of a graph would give all possible one-length walks. N^{th} power of adjacency matrix gave all possible N length walks in the graph. Similarity between two graphs was calculated using direct product graph. Adjacency matrix of direct product graph was found out using Kronecker product. Adjacency matrix of the product graph of graphs $G \ and \ G'$ was calculated using Kronecker product, $A_X = A \otimes A'$. In this work, the similarity between two molecular graphs was found out from the adjacency matrix of direct product graph of edge label filtered adjacency matrix of graphs $G \ and \ G'$.

In case of edge-weighted graphs, edge label filtered adjacency matrix W_x of direct product graph A_X is defined as

$$W_x = \sum_{l=1}^{6} {}^{l}A \otimes {}^{l}A'$$

where ${}^{l}A$ is the edge label filtered adjacency matrix. l is the set of edge labels, $l = \{1,\ 2,\ 3,\ 4,\ 5, 6\}$, where 1 −single bond, 2 −double bond, 3 − triple bond, 4 − $amide$, 5 − aromatic, and 6 −unknown. L is the label matrix, L_{ij} is the label of (v_i, v_j) . Figure 1 explains the label matrix concept. The edge label filtered adjacency matrix ${}^{l}A_{ij} = A_{ij}$ if $L_{ij} = l$ otherwise zero.



0	1	0	0	0
1	0	3	0	0
0	3	0	1	2
0	0	1	0	0
0	0	2	0	0

Figure 1. Label Matrix of Molecular Graph

Similarity between molecular graphs $G \ and \ G'$ was calculated using

$$k_{walk}(G, G') = \frac{1}{|G||G'|} \sum_{k=0}^{n} \lambda^k \, e^T W_x^k \, e \tag{5}$$

where λ is the decaying factor and varies from $0 \leq \lambda \leq 1$ *and* k is the length of walk. Compared to the molecular similarity calculation using adjacency matrix, the edge label filtered adjacency matrix gives more accurate prediction. It helps in differentiating molecules based on the type of bonds labels as wells as the connectivity. For example, it helps to differentiate alkanes from alkenes.

The decision of classifier depends on the value of the decision function

$$f(x) \;= sign(w^T x - \gamma) = sign(\sum_{i=1}^{m} u_i d_i k'_{walk}(G_i, G) - \gamma)$$
$$= sign(\sum_{i \in svindex} u_i d_i k'_{walk}(G_i, G) - \gamma\,) \tag{6}$$

where k'_{walk} is calculated using Eq. 3.

4. Results and Discussion

4.1. Classification Accuracy

This section illustrates the performance of walk kernel for finding the similarity between edge-weighted molecular graphs. Accuracy of the method was tested with PTC and MUTAG data sets. The PTC data set contains a set of 344 compounds classified according to their carcinogenicity [12]. The carcinogenicity was tested in four types of animals, female mouse (FM), female rat (FR), male mouse (MM) and male rat (MR). Active molecules are in +1 class and others are in -1 class. Similarly, MUTAG dataset [13] is another standard dataset consisting of 230 molecules with the information about their mutagenicity in *Salmonella typhimurium* model. In this data set, only 188 molecules are considered as learnable. Out of 188 molecules, there are 125 positive examples and 63 negative examples. Information about the datasets can be found elsewhere [11].

The accuracy of the classifier was calculated using 10-fold cross validation. In this, the datasets are divided into 10 equal parts and the training is done with 9 parts and the testing is done with the 10th part. The same procedure is repeated 10 times by changing the training and testing datasets. Each set of compounds is used once for testing purpose and $(n-1)$ times for training purpose. Using this method, data can be efficiently utilized and the average of the result in each of the individual experiments can be used for evaluating the classifier.

Table 1. Statistics on classification datasets

	FM	FR	MM	MR	MUTAG
+1 class	143	121	129	152	125
-1 class	206	230	207	192	63
max $\lvert G \rvert$	109	109	109	109	40
avg $\lvert G \rvert$	25	25.2	26.1	26.1	31.4

In Table 1, second row gives the number of active compounds ($+1$ class) in PTC dataset and mutagens in MUTAG dataset. Third row gives the number of inactive compounds (-1 class) in PTC dataset and non-mutagens in MUTAG dataset. $\max|G|$ in the fourth row gives the maximum number of atoms in a molecule and $\text{avg}\,|G|$ in the fifth row is the average number of atoms in a molecule. Fourth row gives information about the size of the largest molecular graph in the dataset and fifth row contains average size of the graphs in the data set. Values of C were tried, in which $\mathrm{C} = 0.0011$ gave more accurate result. Result of the proposed method is given in Tables 2 and 3.

Performance evaluation was done by calculating Area Under the Curve (AUC) of ROC curve and accuracy, specificity and sensitivity of the classifier. Specificity and sensitivity are known as classification functions. These are used as statistical measures for analyzing the performance of a binary classifier. Sensitivity or recall rate is the ratio of correctly identified positive candidates and total number of positive candidates in the dataset. Specificity gives the proportion of correctly identified negative candidate and total number of negative candidates in the dataset [4]. These two measures relate to type 1 and type 2 errors. A classifier with 100% specificity and 100% sensitivity cannot be expected.

Table 2. Test result of walk kernel before applying empirical filters

	FM	FR	MM	MR	MUTAG
True Positive	63	59	65	85	111
True Negative	173	169	151	142	42
False Positive	33	61	56	50	21
False Negative	80	62	64	67	14
Specificity	84	73.5	73	74	66.7
Sensitivity	44.1	48.8	50.4	56	88.8
Accuracy	67.7	65	64.3	66	81.4

At the end of phase 1, sets of active and inactive molecules were obtained. The second phase is mainly included in this work for reducing the failure of drug due to lack of drug-likeness of the molecule.

Table 3. Test result of walk kernel after applying the empirical filters

	FM	FR	MM	MR	MUTAG
True Positive	62	59	63	82	110
True Negative	181	173	156	151	53
False Positive	25	57	51	41	10
False Negative	81	62	66	70	15
Specificity	87.9	75.3	75.4	78.7	84.2
Sensitivity	43.4	48.8	48.9	54	88
Accuracy	69.7	66.1	65.2	67.8	86.8

Tables 2 and 3 show the test results of edge-weight based random walk kernel with PTC and MUTAG dataset. Table 2 gives the test result before screening. The effect of filtering can be understood by comparing the last three rows of Tables 2 and 3. An increase in specificity, sensitivity and accuracy of classification can be observed. This increase is mainly due to the reduction in the number of false positives. It reduces the failures in the clinical testing. FromTables 2 and 3, it can be seen that there is a very small reduction in true positives also. It is due to the exceptions to Lipinski's rule [19].

Compared to PTC dataset, MUTAG dataset gives more accurate results. In case of unbalanced dataset, the proportion of positive candidates and negative candidates will not be the same. Global accuracy is not enough for analyzing the performance of a classifier [4] and it may lead to an incorrect conclusion about the performance of the classifier. Receiver Operating Characteristics (ROC) curve is a comparison of two operating characteristics, TPR and FPR characteristics as the criterion changes. It is also known as $sensitivity\ vs\ (1 - specificity)$ plot or $TPR\ vs\ FPR$ plot. The best possible prediction is $100\%\ specificity$ and $100\%\ sensitivity$ and it is represented as the upper left corner in the ROC space. In ROC, evolution of true positive rate versus false positive rate is drawn by varying the function value in the decision function. Using this, different classifiers can be compared. ROC curve for each dataset is given in Figure 2. Depending on the requirement, specificity and sensitivity can be varied by changing the function value in the decision function. AUC is also a parameter for analyzing the performance of the classifier. AUC should be high for good classifier. AUC for the proposed random walk kernel is given in Table 4.

Table 4. AUC for different datasets

	FM	FR	MM	MR	MUTAG
AUC	69%	68.1%	70%	65%	82.45%

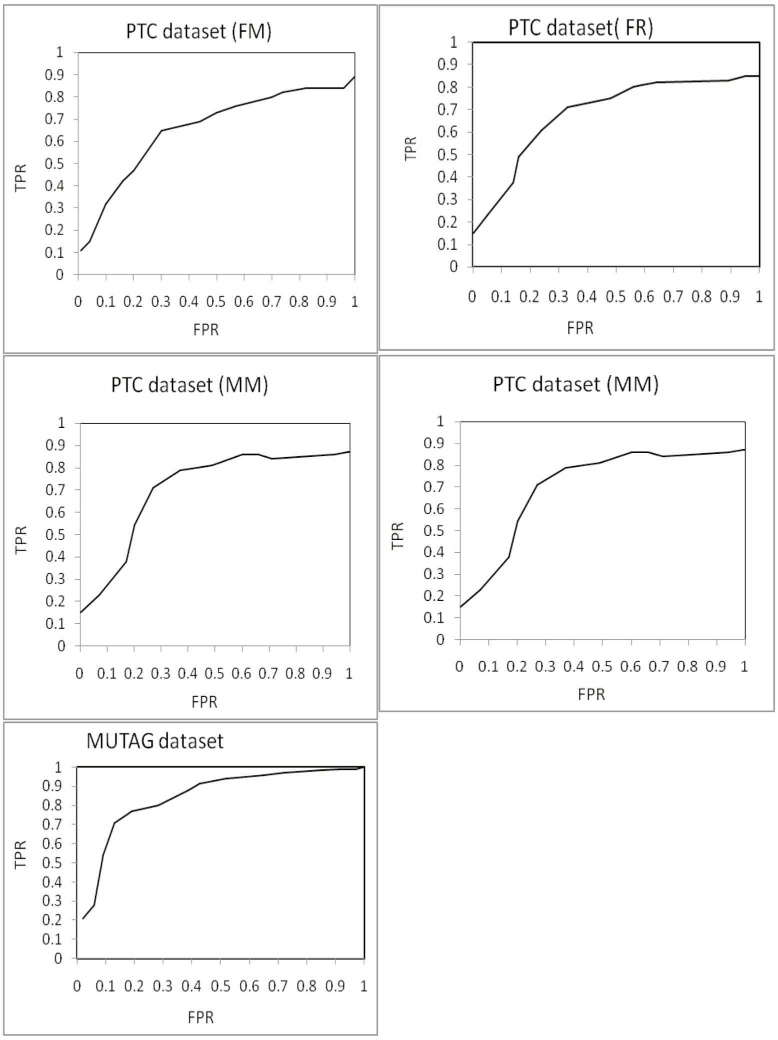


Figure 2. ROC curves for random walk kernel

The points above the diagonal line show classification better than random selection and the points below the diagonal line shows the classification is worse than the random selection. Figure 2 shows that ROC curve for MUTAG dataset is closer to the upper left corner.

Conclusions

Though computationally expensive, results of walk kernel-based classifications are found to be highly reliable. Walk based approach proved to be very efficient for finding the similarity based on linear features. Results indicated that the accuracy was increased when information about the bonds and connectivity together were included for finding the similarity between the molecular graphs. This is because of the fact that type of bond is an important factor for determining the characteristics of a molecule. An initial screening of dataset would help to reduce the number of candidates for the *in vivo* screening. Results have shown that walk kernel helped in avoiding drug failures in the clinical phase. An important challenge in the application of graph kernels in the SAR analysis is the efficient extraction of features and the suitable use of these features for the similarity calculation. In this work, the structural patterns have been used for analyzing the activity of the molecules as the screening using set of traditional descriptors rarely utilize structural patterns. Traditional descriptors have been utilized in the second part of the study. Results have shown that a combination of these two approaches gives very accurate results.

References

[1] Parexel International. “Parexel Biopharmaceutical R&D Statistical Sourcebook 2010/2011.” Waltham, MA: Parexel International, 2010.

[2] Burrill Report. New Estimate of Drug Development Costs Pegs Total at $1.5 Billion. http://ww.burrillreport.com /article - December 07, 2012.

[3] Gasteiger J, Engel T. *Chemoinformatics: A Textbook.* New York: Wiley-VCH; 2003.

[4] Preeja MP, Soman KP. Walk-based graph kernel for drug discovery: A review. *Int J Comput Appl* 2014;94:1-7.

[5] Gärtner T, Flach P, Wrobel S. On graph kernels: Hardness results and efficient alternatives. In: Schölkopf B, Warmuth MK, editors. *Learning Theory and Kernel Machines*, Berlin: Springer 2003, p. 129-143.

[6] Borgwardt KM, Schraudolph NN, Vishwanathan S. Fast computation of graph kernels. In: *Advances in neural information processing system,* Vancouver: 2006, p. 1449-1456.

[7] Tetko IV, Gasteiger J, Todeschini R, Mauri A, Livingstone D, Ertl P, Prokopenko VV. Virtual computational chemistry laboratory – Design and description. *J Comput-Aided Mol Des* 2005;19: 453-63.

[8] VCCLAB, Virtual Computational Chemistry Laboratory, http ://www.vcclab.org , 2005.

[9] Lipinski CA. Drug-like properties and the causes of poor solubility and poor permeability. *J Pharmacol Toxicol Methods* 2000; 44:235-49.

[10] Lipinski CA. Avoiding investment in doomed drugs, is poor solubility an industry wide problem? *Curr Drug Discov* 2001;4:17-9.

[11] Congreve M, Carr R, Murray C, Jhoti H. A ‘rule of three’ for fragment-based lead discovery? *Drug*

Discov Today 2003;8:876-7.

[12] Gleeson MP, Hersey A, Montanari D, Overington J. Probing the links between in vitro potency, ADMET and physicochemical parameters. *Nat Rev Drug Discov* 2011;10:197-208.

[13] Helma C, King RD, Kramer S., Srinivasan A. The predictive toxicology challenge 2000–2001. *Bioinformatics* 2001;17:107-8.

[14] Debnath AK, Lopez de Compadre RL, Debnath G, Shusterman AJ, Hansch C. Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. Correlation with molecular orbital energies and hydrophobicity. *J Med Chem* 1991;34:786-97.

[15] Kashima H, Tsuda K, Inokuchi A. Marginalized kernels between labelled graphs. In: *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)*, Washington DC: 2003, p. 321-288.

[16] Kashima H, Tsuda K., Inokuchi A. . Kernels for graphs. Fast kernels for string and tree matching. In: Schölkopf B, Koji Tsuda K, Vert J-P, editors. *Kernel Methods in Computational Biology*, Cambridge: The MIT Press; 2004, p. 101-113.

[17] Vishwanathan, SVN, Smola, A.J. Fast kernels for string and tree matching. In: Schölkopf B, Koji Tsuda K, Vert J-P, editors. *Kernel Methods in Computational Biology*, Cambridge: The MIT Press; 2004, p. 113-130.

[18] Borgwardt, KM. Graph Kernels. PhD thesis, Ludwig-Maximilians University, Munich, 2007.

[19] Thomas HK., Arkadius P, Zheng Y. A practical view of ‘druggability’ *Curr Opin Chem Biol* 2006;10:357–61.