ISSN No. 0976-5697

Volume 3, No. 3, May-June 2012

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Molecular Similarity Based on Functional Groups: Subtree Kernels for Virtual Screening in Drug Discovery

M. P. Preeja*
Biotechnology Engineering
Sahrdaya College of Engineering and Technology
Thrissur, India
preeja_mp@yahoo.com

K. P. Soman
Center for Excellence in Computational Engineering and Networking
Amrita Vishwa Vidyapeetham, Coimbatore, India
kp_soman@amrita.edu

Abstract: Current developments in computer-aided drug design (CADD) reduces the time and cost of drug discovery process. Using various computational techniques, large number of compounds in the chemical database are analyzed and screened. Virtual screening can be classified into structure based and ligand based methods. The amount of information required for structure based virtual screening is too high when compared to ligand based method. In ligand based methods, classification of active and inactive drugs can be done using Hidden Markov Model, Support Vector Machine (SVM), Clustering etc. SVM is widely used nowadays. Classification using SVM and graph kernel function is introduced recently and shows good accuracy over the other existing methods. Different graph kernel functions are used for finding similarity measure in SVM. Here, we propose a new method for finding the similarity based on the functional groups present in the molecules. The accuracy is compared with that of existing graph kernel methods using standard data sets (PTC and MUTAG).

Keywords: drug discovery; virtual screening; SVM; molecular graph similarity functional group; tree kernel.

I. INTRODUCTION

Computational techniques are used in the early stages of drug discovery. Computational techniques reduce the time and cost of drug discovery process. Mainly it is applicable in combinatorial chemistry, virtual screening, quantitative structure-activity relationship (QSAR) and drug lead optimization. Natural extracts are the main source of drugs. Efficient drugs for each disease are the correct combination of these natural extracts. Combinatorial chemistry produces millions of compounds. Inactive compounds from this pool of compounds are filtered using virtual screening [1]-[4]. Efficient virtual screening reduces the number of compounds for the clinical test.

Virtual screening methods are classified into two categories- structure based and ligand based. Structural based virtual screening is based on the ligand-target molecule binding. It requires information about both the target and ligand. Ligand based virtual screening is based on the structural similarity of the ligand molecules. In the initial stage of drug discovery, ligand based methods are commonly used because of lack of information about target. In ligand based methods, filtering is mainly based on the absorption, distribution, metabolism, excretion and toxicity (ADMET) properties of the molecule [5]. In drug discovery, 50-60% failure is because of the poor ADMET properties of the drug molecule. ADMET properties of the molecule can be analyzed based on topological and structural similarity [2]-[4].

Ligand based virtual screening techniques are mainly classified into five - small molecule similarity, pharmacophore based search, descriptor based search, recursive partitioning, and graph based similarity search [6] among which graph based similarity methods are widely used because graphs can store information about atoms in the molecule and the connections between these atoms [7]. Graph based virtual screening is based on the structural activity relationships. The characteristics of molecules will be similar if their structures are similar [8]. Graph comparison requires comparison of nodes and edges connecting these nodes. Two graphs are similar if their substructures are similar. Different substructures used for the graph comparison are walks, paths, cycles, sub-trees and subgraphs [9]-[15]. In machine learning, classification of molecules is based on the molecular similarity measurement; molecular similarity can be measured in different ways [7]-[18]. Methods for similarity measurement are different for numerical and structured data. Structured data requires preprocessing for finding the similarity [7]. In graph kernel, substructures are the features for the comparison. Based on the substructures used, graph kernels can be classified into walk kernel, marginalized kernel, shortest path kernel, sub- tree kernel etc [11]. Sub-tree kernels use non linear features for comparison and walk kernels use linear features for comparison [11].

In this paper, we propose sub-tree kernels for finding the similarity based on the functional groups present in the molecules. In [5] and [7], authors show that most of the ADMET properties are based on the presence of functional groups in that molecule. Reminder of this paper is organized as follows: In section II, different types of molecular graphs are recalled; SVM and graph kernel based classification is explained in section III; sub-tree kernels for functional group similarity is presented in section IV; results and the comparison of sub-tree kernels is explained in section V.

II. REPRESNTATION OF MOLECULE

Chemical compounds are represented in different ways based on the availability of information and requirements for each application. One dimensional (ID) representation of the molecule gives information about the composition. Two dimensional (2D) representations give the structure of the molecules. In most of the pattern recognition methods, 2D representation of the molecules is used. Three dimensional (3D) representation of the molecule gives the shape of the molecule. Ligand based virtual screening requires ID and 2D representation while structure based virtual screening requires 3D representation.

Graph kernel requires 2D representation of molecules. Different types of molecular graphs are used for finding the similarity - unlabelled and labeled, directed and undirected, weighted and unweighted etc. Fig. 1 is an example of labeled undirected molecular graph.

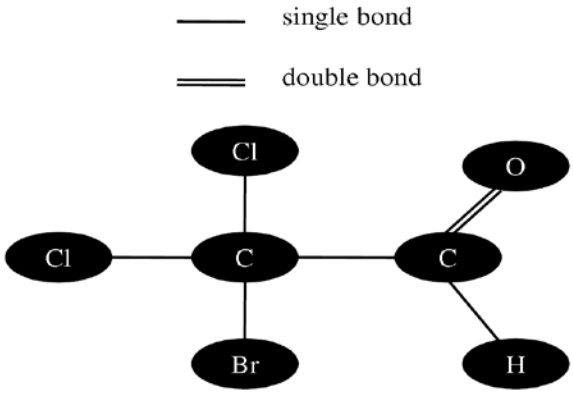

Figure 1. Labeled undirected molecular graph

A. *Molecular Graph:*

Molecular graph $G = (V, E)$ consist of finite sequence of vertices (atoms) and edges (bonds between atoms). $L_V(G)$ is the set of vertex labels and $L_E(G)$ is the set of edge labels. Adjacency matrix $A(G)$ gives information about the connectivity.

$$A_{ij}(G) = \begin{cases} 1 \; if \;\; (v_i, v_j) \in E \\ 0 \;\; otherwise \end{cases}$$

III. METHODS

A. *SVM for Virtual Screening:*

In virtual screening, non drug molecules are filtered from the pool of compounds. SVM [19] is used to classify active (drug molecule) and inactive (non drug molecule) compounds. In binary classification, training data is of the form (x_i, y_i) $i = 1,...m$ where m is the number of training data. x_i is the feature vector and y_i is the corresponding output. In binary classification, $y_i \in \{-1, +1\}$. Finding the coefficient for the classifier is a convex optimization problem. Soft margin SVM finds the classifier by solving the following convex optimization problem

$$\min_{w,\gamma,\xi} \quad \frac{w^T w}{2} + C \sum_{i=1}^{N} \xi_i \qquad (1)$$

Subject to

$$y_i(w^T x - \gamma) + \xi_i, \; 1 \le i \le N$$

$$\xi_i \ge 0, 1 \le i \le N$$

where C is the controlling parameter.

$$If \;\; C = high\ value(\infty) \;\; \xi_i \; is\ not\ allowed$$

$$otherwise \; \xi_i \; is\ allowed$$

$w = \sum_{i=1}^{N} u_i y_i x_i$ *where* u_i is the Lagrangian multiplier and $\gamma = \sum_{i \in SV} u_i y_i x_i$ *where sv* is the set of support vectors.

The decision function for new data is

$$f(x) = sign \; (\sum_{i \in SV} u_i y_i \langle x, x_i \rangle) \qquad (2)$$

where $\langle x, x_i \rangle$ is the linear kernel function or similarity between x *and* x_i.

Non-linearly separable data kernel function is $\langle \phi(x), \phi(x_i) \rangle$ *and* $\langle x, x_i \rangle$ *is for linearly seperable data.*

B. *Graph Kernel:*

Finding similarity between structured data (eg. molecular grap hs) requires substructures as features. Different substructures of graphs are walks, paths, subtrees, subgraphs etc. [11]. Walks and paths are the linear features. Subtrees are non linear features and these include information about the functional groups present in the molecule.

Graph kernel between two graphs G_1 *and* G_2 is defined as

$$k(G_1, G_2) = \sum_{i \in substruct_1} \sum_{j \in substruct_2} k(substruct_i, substruct_j) \qquad (3)$$

where $substruct_1$ *and* $substruct_2$ are the set of substructures of graph G_1 *and* G_2.

In [11], the authors introduced subtree kernels for virtual screening in drug discovery. The authors proposed a subtree kernel for avoiding the limitation of linear features like walk and path and proposed that linear features cannot include full information about functional units. These features give only partial information about the functional units. In [20], the authors made modifications for avoiding tottering effect and used atom label enrichment by using Morgan indexing for vertex labels. In [21], the authors proposed fast computation subtree kernels. The authors proposed methods for reducing the kernel computing complexity upto $O(mh)$ where m is the number of nodes and h is the height of the tree.

Subtree kernel is defined as

$$k(G_1, G_2) = \sum_{i \in st_1} \sum_{j \in st_2} k(st_i, st_j) \quad (4)$$

where st_1 *and* st_2 are the sets of subtrees of graphs G_1 *and* G_2.

In this paper, we propose efficient subtree kernels for finding the similarity between molecular graphs. We compare all the possible subtrees of depth two. These subtrees give information about the functional groups. In [11] and [20], the authors considered subtrees with different depths.

Subtree decomposition is done by using adjacency matrix (A). i^{th} row ($A_{ij}, j = 1....n$) in the adjacency matrix represents i^{th} atom and its neighbors. It represents subtree with parent as i^{th} atom and children as its neighbors.

Similarity between molecular graphs G_1 *and* G_2

$$k_{subtree}(G_1, G_2) = \sum_{i=1}^{n} \sum_{j=1}^{m} k(st_i, st_j) \quad (5)$$

where st_i is the subtree from graph G_1 and st_j is the sub-tree from graph G_2 . m *and* n are the numbers of atoms in graph G_1 *and* G_2.

Similarity between subtrees (Fig. 2) st_i and st_j is calculated by

a. Comparing the labels of subtrees
b. Comparison based on the number of branches of subtree

$$k(st_i, st_j) = \begin{cases} \delta_{11} + \sum_{s=2}^{k} \sum_{t=2}^{h} \delta_{st} \\ 0 \ \ if \ \ k \neq h \end{cases} \quad (6)$$

Where $\delta_{st} = \begin{cases} 1 \ if \ \ label(st_{is}) = label(st_{jt}) \\ 0 \ otherwise \end{cases}$ and

k = number of nodes in st_i

h = number of nodes in st_j

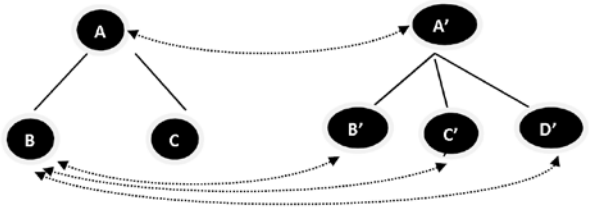

Figure 2. Subtree comparison

Normalization of the kernel is done by using

$$k'_{subtree}(G_1, G_2) = \frac{k_{subtree}(G_1, G_2)}{\sqrt{k_{subtree}(G_1, G_1) * k_{subtree}(G_2, G_2)}} \quad (7)$$

IV. RESULTS AND DISCUSSION

In this section, we illustrate the performance of subtree kernel for finding the similarity between molecular graphs based on the presence of functional group. Accuracy of the method is tested with Predictive Toxicology Challenge (PTC) and MUTAG data sets. PTC data set contain 344 compounds and its activity is 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 in -1 class. MUTAG is another standard dataset consisting of 188 compounds classified as mutagens or non-mutagens. Mutagens are denoted as +1 class and non mutagens are denoted as -1 class. Information about the data is given in Table 1.

Table : 1 Dataset for classification

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

In Table I, first row gives number of +1 class data points, active compounds in PTC dataset and mutagens in MUTAG dataset. Second row gives number of inactive compounds in PTC dataset and non mutagens in MUTAG dataset. $\max|G|$ in the third row gives maximum number of atoms in a molecule and $avg|G|$ in the fourth row is the average number of atoms in a molecule. Third and fourth row gives information about the size of graphs we need to classify and time complexity for the comparison.

Performance evaluation of the proposed method is done based on the following criteria.

$$\text{Specificity} = \frac{TN}{TN + FP} \times 100$$

$$Sensitivity = \frac{TP}{TP + FN} \times 100$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100$$

Where TP and TN, true positive and true negative respectively, represent correctly classified compounds. FN and FP, false negative and false positive, represent misclassified compounds. Sensitivity is the correct classification rate of +1 class and specificity is the correct classification rate of -1 class.

We evaluate the method using 10-fold cross validation. This allows the maximum use of available dataset. Result of the proposed method is given in Table II and III. Table II gives information about the training phase and Table III shows the accuracy of our model in the testing phase. Table I shows PTC dataset contain more number of -1 class molecules while MUTAG dataset contains more number of +1class molecules. Eighth and ninth rows in Table II shows

PTC dataset specificity is high compared to sensitivity because more number of -1 class data points for the training while in MUTAG specificity is low compared to sensitivity due to the less number of -1 class data for the training. The accuracy of the proposed method is good compared to other existing method .In this model similarity is based on the comparison of subtrees of depth two. Result shows using subtrees of depth two we can include most of the information about functional groups.

Table:2 Training Results

	FM	FR	MM	MR	MUTAG
Best C	0.56	0.62	0.59	0.57	0.96
Number of support Vectors	163	178	148	189	93
True Positive	140	115	120	141	123
True Negative	200	220	200	185	60
False Positive	3	6	9	11	2
False Negative	6	10	7	7	3
Specificity	98	97.34	95.6	95.3	96.71
Sensitivity	95.6	92	94.48	95.33	97.6
Error	2.5	4.5	4.78	4.6	2.6
Accuracy	97.42	95.45	95.23	95.34	97.34

Table: 3 Test Results

	FM	FR	MM	MR	MUTAG
Best C	0.56	0.62	0.59	0.57	0.96
Number of support Vectors	163	178	148	189	93
True Positive	63	59	65	85	109
True Negative	173	169	151	142	39
False Positive	33	61	56	50	24
False Negative	80	62	64	72	16
Specificity	83.98	73.47	72.94	73.95	61.90
Sensitivity	44.05	48.76	50.38	54.14	87.2
Error	32.37	35.05	35.72	34.96	21.27
Accuracy	67.63	64.95	64.28	65.04	78.73

V. CONCLUSION

In this paper, we proposed new subtree kernel for virtual screening. Results show that nonlinear feature gives more accurate prediction than linear features like random walk. Subtree kernel with depth two capture most of the information about functional groups. Most of the drug like properties is because of the presence of specific functional group. This kernel can be useful in other applications in information technology.

VI. ACKNOWLEDGEMENT

We would like to thank Andreas Bender, Lecturer in Molecular Informatics, Unilever Centre for Molecular Science Informatics at the University of Cambridge, for his help and guidance in carrying out this research.

VII. REFERENCES

[1] Terrett, N. K.; Gardner, M.; Gordon, D. W.; Kobylecki, R. J.; Steele, J. Drug discovery by combinatorial chemistry - the development of a novel method for the rapid synthesis of single compounds. *Chem. Eur. J.* **1997**, *3*, 1917-1920.

[2] Tingjun HOU and Xiaojie Xu , "Recent Development and Application of Virtual Screening in Drug Discovery: An Overview," Current Pharmaceutical Design, 2004, *10*, 1011-1033

[3] Vivek Vyas , Anurekha Jain, Avijeet Jain, Arun Gupta, "Virtual Screening: A Fast Tool for Drug Design, " Sci Pharm., pp. 333-360, vol. 76, 2008.

[4] Markus H, J Seifert and Martin Lang, "Essential Factors for Successful *Virtual Screening*", Mini-reveiws in Medicinal Chemistry, 2007, 7, 63-72.

[5] Dennis A. Smith, "Metabolism, Pharmacokinetics and toxity of Functional Groups:Impact of Chemical building blocks on ADMET," RSE drug Discovery Serious No.1.

[6] A. Srinivas Reddy, S. Priyadarshini Pati, P. Praveen Kumar, H.N.Pradeep and G. Narahari Sastry, "Virtual Screening in Drug Discovery – A Computational Perspective," Current Protein and Peptide Science, 2007, 8, 329-351 .

[7] C. A. Azencott, A. Ksikes, S. J. Swamidass,J. H. Chen, L. Ralaivola, and P. Baldi, "One- to fourdimensional kernels for virtual screening and the prediction of physical, chemical, and biological properties," J Chem Inf Model, vol.47 no. 03 pp. 965-974, 2007.

[8] S. Kramer, E. Frank, and C. Helma, "Fragment generation and support vector machines for inducing SARs," SAR QSAR Environ Res., vol. 13, no. 05, pp. 509-523, July 2002.

[9] Helma, T. Cramer, S. Kramer, and L. De Raedt, "Data mining and machine learning techniques for the identification of mutagenicity inducing substructures and structure activity relationships of noncongeneric compounds," J Chem Inf Comput Sci., vol. 44, no. 04, pp. 1402-1411, 2004.

[10] R. Burbidge, M. Trotter, B. Buxton, and S. Holden, "Drug design by machine learning: support vector machines for pharmaceutical data analysis," Comput. Chem., vol 26, no 014-15, December 2001.

[11] J. Ramon and T. Gartner, "Expressivity versus effciency of graph kernels," Proceedings of the First International Workshop on Mining Graphs, Trees and Sequences, pp. 65-74, 2003.

[12] P. Mahe, N. Ueda, T. Akutsu, J.-L. Perret, and J.-P. Vert, "Graph kernels for molecular structure-activity relationship analysis with support vector machines," J Chem Inf Model, vol. 44 no. 04 , pp. 939-951, 2005.

[13] D. Haussler, "Convolutional kernels on discrete structures", Technical Report UCSC-CRL-99- 10, Computer Science Department, UC Santa Cruz, 1999.

[14] H. Kashima, K. Tsuda and A. Inokuchi, "Marginalized Kernels between Labeled Graphs," Proceedings of the Twentieth International Conference on Machine Learning, pp. 321-328, AAAI Press, 2003.

[15] T. Gartner, P. Flach, and S.Wrobel, "On graph kernels: hardness results and efficient alternatives," Proceedings of the Sixteenth Annual Conference on Computational Learning Theory and the Seventh Annual Workshop on Kernel Machines, vol. 2777 of Lecture Notes in Computer Science, pp. 129-143, July 2003.

[16] K. M. Borgwardt and H.-P. Kriegel, "Shortest-path kernels on graphs,"Proc. 5th IEEE Int. Conf. on Data Mining, pp. 74-81,2005.

[17] P. Mahe and J.P. Vert, "Graph kernels based on tree patterns for molecules," Technical Report ccsd-00095488, HAL,September 2006.

[18] Iain Martin, "The Influence of PhysicochemicalProperties on ADME," Physchem Forum 2.

[19] Cristianini, N. and J. Shawe-Taylor: 2000, An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge University Press.

[20] Pierre Mahé and Jean-Philippe Vert, "Graph Kernels based on tree patterns for molecules," Journal Machine Learning archive," Volume 75 Issue 1, April 2009