International Journal for Computational Biology (IJCB)
Vol.1, No.1, 2012, Page. 56-62
ISSN: 2278-8115

Application of Support Vector Machines in Virtual Screening

Soumi Sengupta, Sanghamitra Bandyopadhyay*
Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India-700108

Article Info

Article history:

Received Jan 10th, 2012
Revised Feb 5th, 2012
Accepted Feb 13th, 2012

Keyword:

Drug Design
Virtual Screening
Quantitative Structure Activity Relationship
Support Vector Machines.

ABSTRACT

Traditionally drug discovery has been a labor intensive effort, since it is difficult to identify a possible drug candidate from an extremely large small molecule library for any given target. Most of the small molecules fail to show any activity against the target because of electrochemical, structural and other incompatibilities. Virtual screening is an in-silico approach to identify drug candidates which are unlikely to show any activity against a given target, thus reducing an enormous amount of experimentation which is most likely to end up as failures. Important approaches in virtual screening have been through docking studies and using classification techniques. Support vector machines based classifiers, based on the principles of statistical learning theory have found several applications in virtual screening. In this paper, first the theory and main principles of SVM are briefly outlined. Thereafter a few successful applications of SVM in virtual screening have been discussed. It further underlines the pitfalls of the existing approaches and highlights the area which needs further contribution to improve the state of the art for application of SVM in virtual screening.

Corresponding Author:

Sanghamitra Bandyopadhyay,
Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India
Email: sanghami@isical.ac.in

How to Cite:

Soumi Sengupta et. al. Application of SupporVector Machines inVirtual Screening. IJCB. 2012; Volume 1 (Issue 1): Page 56-62.

1. INTRODUCTION

Rational drug design is a focused approach to aid traditional drug discovery to reduce experimental cost and time. This basically involves identification or creation of candidate drug like molecule using the information about the structure of a drug receptor or one of its natural ligands. It includes four essential steps as shown in Fig. 1.

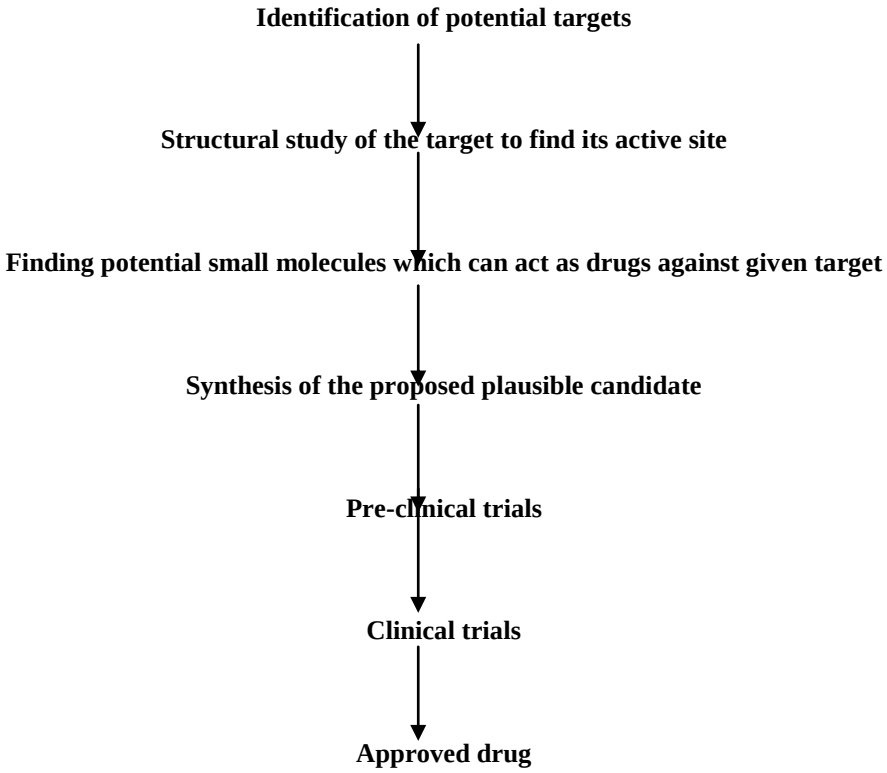

Fig. 1 Essential steps of drug discovery procedure

Virtual screening is basically an approach to search the whole known chemical space with the aid of computational techniques. It strives to find novel molecular scaffold which can act as drug against a particular given target protein. It focuses on defining the criteria for filtering the large chemical databases and applying them to find small molecules with desired properties which can bind to a given drug target and act as a drug. Several approaches used for virtual screening includes Hansch analysis, Free-Wilson analysis, Hansch Free-Wilson mixed approach[1], [2], active site interactions[3] or de novo models[4], [5], 3D-pharmacophore based design[6], comparative molecular field analysis CoMFA[7], [8] and molecular docking[9], [10]. Quantitative structure activity relationship, QSAR [11], [2] is an approach to rational drug design that is based on the concept that the structure and biological properties of a molecule are interrelated. It includes Hansch and Free-Wilson analyses. Hansch analysis correlates physicochemical properties to biological activity of a molecule. Free-Wilson is a numerical method which directly relates structural features with biological properties. Both approaches are closely interrelated, theoretically as well as in their practical applications. Thus, both the approaches are often used combinatorially where, biological activity due to certain structural modifications are calculated using Free-Wilson type parameters and physicochemical parameters are used to describe the effect of some other substituent on the biological activity[12]. Basically, in QSAR approach, a molecular structure is represented using the computable molecular descriptors (broadly hydrophobic, electronic and steric) to compute the biological activity with the equations relating them. These equations are used to calculate the biological activity of different synthesized and predicted molecules against a given target to ensure their effectiveness as a drug.

Pharmacophore can be defined as “a set of structural features in a molecule that is recognized at a receptor site and is responsible for that molecule’s biological activity” [13]. Pharmacophore modeling involves the following steps that are iteratively performed until a desired result is obtained: Identification of structural and chemical features common in biologically active molecules, measurement of three dimensional orientation of the common features, defining a pharmacophore, validation of the pharmacophore to be harbored by active compounds and not by any inactive compound and refining of the pharmacophore model by applying the same to compounds with known functionality. This approach is more useful when the structural information of the

target protein in unknown. In [14] and [15] pharmacophore modeling has been used as a tool to screen large libraries to find inhibitors of given target.

The principal concept underlying Comparative Molecular Field Analysis (CoMFA), is that the difference in biological activity of a molecule is due to the changes in shapes and strength of its non-covalent interaction (steric and electrostatic) to its surroundings. For CoMFA analysis, a set of molecules is required. Each conformation of these molecules is considered to be the active structure, and is placed on the cubic grid to calculate its biological activity by finding its surrounding molecular field using appropriate probe. An article enunciating a study on epothilones using CoMFA is discussed in [8]. Docking is also an efficient approach to rational drug design which tries to "predict the structure and binding free energy of a ligand receptor complex given only the structure of the free ligand and receptor" [1]. An automated docking program, DOCK, was proposed in [3]. A docking problem can be broken to basic three steps as shown in Fig. 2: (i) defining the potential drug target and identification of its active site to which the drug molecule must bind, (ii) modeling a drug like small molecule and study of its interaction with the receptor protein, and (iii) performing the conformational and orientation search to find low energy states of the system that can correlate to the original binding model.

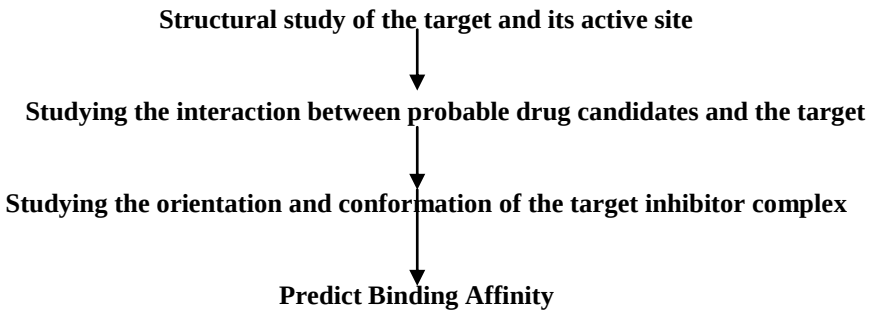

Fig. 2 Steps involved in docking studies

Active site of a protein is a localized combination of amino acid groups within its tertiary (3D) or quaternary structure that are capable of interacting with a chemically specific substrate which provides the protein with its biological activity. Virtual screening involving active site interactions can be categorized into two groups. The first approach strives to find molecules interacting optimally to the active site by searching different databases of known small drug like molecules and second, involves building de novo models of ligands complementary to a given active site. De novo design can be done in two ways: Outside-in and Inside-out approaches. In Outside-in approach the binding site is analyzed to determine where specific functional groups might bind properly so that these functional groups can be connected to build a real molecule. On the other hand, Inside-out approach involves growing a ligand molecule within the given active site using an appropriate search algorithm where each proposed ligand is evaluated using an energy function [4], [16], [17].

Several computational concepts have been facilitated for the above mentioned approaches of virtual screening. In the present article we would emphasize on the use of support vector machines (SVM) for these virtual screening approaches.

2. SUPPORT VECTOR MACHINES

Support vector machines was invented by Vapnik et.al.,in 1979 [18], [19], [20]. It is a machine learning technique to facilitate classification. Two basic concepts essential for support vector machines are a maximal margin classifier and a kernel function. The former is responsible for the construction of a separating hyperplane so that the distance between the different classes is maximized. The latter is used to map the data in a new space where they are separable.

Firstly the SVM is trained using a learning data set. This data set necessarily contains data divided in two classes. When these training data are linearly separable the algorithm learns to construct the unique hyperplane having the maximal margin (δ) separating the training objects into two classes as show in Fig. 3a. However, the data can also be linearly inseparable. In such a scenario the algorithm projects the input data vectors to a higher dimensional feature space using kernel functions [20], [21]. Thereafter these projections of the data are classified by constructing a hyperplane as shown in Fig. 3b.

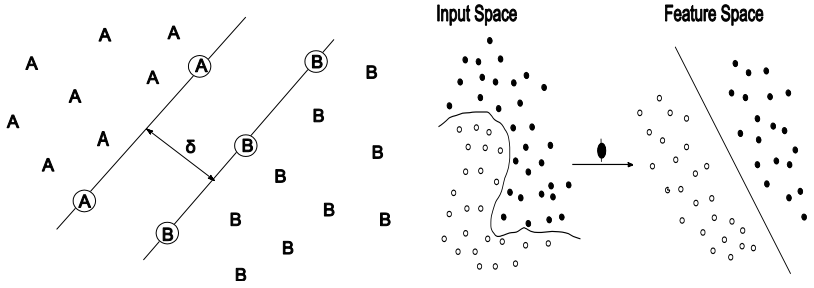

Fig. 3 (a) Maximum separation hyperplane for linearly separable data, (b) Linear separation in feature space for linearly inseparable data.

There are several kernel functions developed for this purpose. A recent review in 2007 delineated the use of kernel functions from the perspective of biological and chemical applications. This article vividly discusses the necessity and usage of different kernel functions. It states that if a data set can be linearly classified then non-linear relations should be avoided for its classification. It also shows that overfitting of the training data can occur when more complex kernel functions are used. For solving such a problem the author recommends comparing SVM models based on non-linear kernel functions with SVM models obtained with a linear kernel since the separating hypersurface may be almost linear. In similar fashion the author have discussed and explained several rules for kernel selection and comparison of the results of different kernel functions.

3. APPLICATION OF SVM IN VIRTUAL SCREENING

Virtual screening basically involves examination of many molecules to find active molecules against a given target. This is generally done with the knowledge of the known inhibitors of the given target. If the molecular properties of a set of molecule which is active against a target of interest are known then the vast chemical space can be searched to fetch such molecules which possess the same properties. If a classifier is trained with the molecular properties of the known inhibitors of a given target then the classifier can predict or classify other molecules from the chemical databases to active or inactive against the same target. Therefore the problem of virtual screening can be well framed as a classification problem. The ability of SVM to successfully classify linearly and non linearly separable data has made its application popular in drug design, virtual screening and combinatorial chemistry [22], [23], [24], [25], [26], [27].

In [28] the authors have reported results of such an analysis using SVM. Here molecules in five data sets were ranked according to their activity against given targets. This approach uses Gaussian kernel and those molecular descriptors for which the values of active and inactive molecules lie in different and distinct ranges. SVM has also been applied for activity prediction of small molecules other than their classification and ranking.

Yao et. al., had proposed an SVM based technique to predict the activity of cyclooxygenase 2 (COX-2) inhibitors [29]. The kernel function used for the study was radial basis function. The structure of the molecules was described using several molecular descriptors. These include one constitutional descriptor, one geometrical descriptor, one topological index, one electrostatic descriptor, and two quantum chemical descriptors. Another approach in [30] elucidates the application of SVM for drug-likeness and agrochemical-likeness prediction to aid virtual screening. It also successfully predicts the activity of several Carbonic Anhydrase II (CA II) enzyme inhibitors. The drug and agrolikeness descriptors used for the study are as follows: molecular weight, fractional absorption, log of 1-octanol/water partition coefficient at pH 7.4, log of 1-octanol/water partition coefficient (neutral form), log of water solubility (g/mL) at pH 7.4, number of hydrogen bond donors, number of hydrogen bond acceptors, number of rotatable bonds, molecular radius of gyration, atomic polarizability, dipole moment, and set of Jurs descriptors. This study demonstrated that SVM with radial basis function kernel outperforms SVM with linear kernel, multilayer perceptron, modular feed forward network and generalized feed forward network. Though, the usage of drug and agro-likeness descriptors help in predicting drug likeness and agrochemical-likeness but the authors have stated that usage of 2D and 3D descriptors will improve prediction rate. SVM models are also used for feature selection. In this case the algorithm selects the features most essential for the prediction of activity/ druggability/ toxicity/ any other property of a set of molecules. Byvatov and Schneider [31] discuss a work that uses SVM for selecting relevant

molecular descriptors from a trained classifier which could be important for better understanding of ligand-receptor interactions. In Muller.et.al.,[32] another SVM based technique for drug-likeness prediction is presented. In these types of studies where feature selection and prediction are done together, the classifier is trained first using all the features available and then a feature selection algorithm is employed to discard the least important features. But the main drawback of this approach is that the classifier assumes a statistical distribution the data which may/ may not be a good approximation. Also the feature selection becomes dependent on the classification model. Therefore, these issues must always be handled while amalgamating feature selection with a prediction/ classification model.

In [33] an in silico chemogenomics approach for virtual screening of G-protein coupled receptors has been discussed. This is a SVM based approach that uses a flexible framework to incorporate various information sources on the biological space of targets and on the chemical space of small molecules. The article also investigates the usage of 2D and 3D descriptors for small molecules to gauge the prediction accuracies. It shows that the inclusion of the information about the known hierarchical classification of the target family and the interacting amino acids in the binding pockets of the target under consideration significantly improves the prediction accuracy of the proposed model.

Recently a few works delineates a shift from the usual paradigm of virtual screening. These approaches use the pair potentials of active and inactive molecules instead of only their molecular descriptors to train the SVM. Li et. al., [34] use the same approach to identify EGFR kinase inhibitors. The positive data of the training set containes the pair potentials obtained from three-dimensional structures of protein-ligand complexes present in PDB. The negative data set constituted of the computed pair potentials of protein-ligand complexes obtained by docking a set of randomly chosen 10,000 molecules to EGFR. To further fine tune their results Li et. al., [34] compared the binding profile of the predicted active molecules to the binding profile of the known EGFR inhibitor, erlotinib. Here, the binding profile is referred to as the off-targets of a molecule occurring in multiple signalling pathways. This profile contains the binding affinity of a molecule to different targets spanning over several signaling pathways. Using this approach, Li et. al., were able to identify three new inhibitors of EGFR. A similar and more improved approach was proposed in [35]. The authors defined new pair potentials based on 2018 protein-ligand complexes. The negative data set in this case was more intelligently designed. It contained the computed pair potentials of the protein-ligand complexes where, the ligands were inactive molecules or decoys obtained from Directory of Useful Decoys (DUD). The rationale behind designing such a negative data set was to ensure that the SVM can efficiently distinguish between molecules that acquire the binding modes of active molecule and not the decoys. These are new approaches to virtual screening using SVM which produces promising results but, these approaches are target as well as inhibitor type specific. Therefore, virtual screening for different targets requires selection of new set of pair potentials and construction of binding profiles. Moreover, amongst these predicted molecules only very few are found to be really active against the given target when tested experimentally. Though SVM has lower prediction error in comparison to other classifiers but still it needs more endeavor to make more relevant biological predictions.

There have also been endeavors in developing new kernel functions which could be more helpful for biological predictions. Mostly SVM applications, which are used for virtual screening uses the radial basis function. In [36] a novel graph alignment kernel function is proposed, which is used for virtual screening. This method uses graphs to represent molecules. Then it applies wavelet analysis on these graphs to capture the local topologies of these molecules. The features generated in this way have been further used by the novel graph alignment kernel function to build SVM models for virtual screening. A similar approach is also discussed in [37] which is an extension or improvement of the former work. Here the kernel function is a bit improved and is termed as graph assignment kernel.

4. CONCLUSION

The applications discussed in this paper enunciate the effectiveness of the machine learning approaches in virtual screening. It has some definitive advantages over random selection. Therefore, it is evident that virtual screening can make important contributions to the drug discovery process. The application of machine learning is particularly beneficial when the objective is to reduce a large data set to a smaller chemical library. But, it must is noticed that the efficiency of these methods completely depends on the quality of the data set being used. Feature selection is also important for predictive model building. When feature selection and training of the model occurs simultaneously it should be taken care that the statistical distribution of the data has been chosen appropriately.

Most studies report that the SVM performs better than other machine learning approaches. But, still performance of these algorithms needs to be improved. The prediction quality of SVM can often be improved by adjusting its parameters to the particular problem. Specifically the kernel function selection for a particular

problem is difficult. Therefore, further studies are required to be made for improving the performance of SVM in virtual screening particularly.

REFERENCES

[1] Blaney JM and Dixion JS, "On the Information Content of 2D and 3D Descriptors for QSAR", Perspect Drug Discov.Des., 1: 301–319 (1993).
[2] Ghosal N and Mukherjee PK, "3D QSAR of N-substituted 4-amino-3,3-dialkyl-2(3H)-furanone GABA Receptor Modulators Using Molecular Field Analysis and Receptor Surface Modeling Study", Bioinorg. Med. Chem. Lett., 14:103–109 (2004).
[3] Kuntz I D, Blaney EC, Oatley SJ, Langridge R, and Ferrin TE, "A Geometric Approach to Macromolecule-Ligand Interactions", J. Mol. Biol., 161: 269–288 (1982).
[4] Bandyopadhyay S, Bagchi A, and Maulik U, "Active Site Driven Ligand Design: An Evolutionary Approach", Journal of Bioinformatics and Computational Biology, 3:1053–1070 (2005).
[5] Goh G and Foster JA, Evolving molecules for drug design using genetic algorithm via Molecular Tree in Int Conf. Genet. Evol.Comput 2000, 27–33.
[6] Jones G, Willett P, and Glen RC, "A genetic algorithm for flexible molecular overlay and pharmacophore elucidation", J. Comput. Aided. Mol. Des., 9: 532–549 (1995).
[7] Kimura T, Hasegawa K, and Funatsu K, "GA Strategy for Variable Selection in QSAR Studies: GA-Based Region Selection for CoMFA Modeling", Journal of Chemical Information and Computer Sciences, 38: 276–282 (1998).
[8] Lee KW and Briggs JM, "Comparative Molecular Field Analysis (CoMFA) study of Epothilones – Tubulin Depolymerization Inhibitors: Pharmacophore Development
Using 3DQSAR Methods", J. Comput. Chem., 15: 41–55 (2001).
[9] Pegg SC, Haresco JJ, and Kuntz ID, "A Genetic Algorithm for Structure-based De Novo Design", J. Comput. Aided. Mol. Des., 15: 911–933 (2001).
[10] Oshiro CM, Kuntz ID, and Dixion JS, "Flexible ligand docking using a genetic algorithm", J. Comput. Aided. Mol. Des., 8: 565–582 (1994).
[11] Oprea TI, "On the Information Content of 2D and 3D Descriptors for QSAR", J. Braz. Chem. Soc., 13: 811–815 (2002).
[12] Kubinyi H, "Free Wilson Analysis. Theory, Applications and its Relationship to Hansch Analysis", Quantitative Structure-Activity Relationships, 7: 121–133 (1988).
[13] Guner O, Pharmacophore Perception, Development, and use in Drug Design,
International University Line : La Jolla, 254—268; pp 254—268.
[14] Hecker EA, Duraiswami C, Andrea TA,
and Diller DJ, "Use of catalyst pharmacophore models for screening of large combinatorial libraries.", J ChemInfComputSci, 42: 1204–1211 (2002).
[15] Liu F, You Qi-Dong, Chen Ya-Dong, "Pharmacophore identification of ksp inhibitors", Bioorganic & Medicinal Chemistry Letters, 17: 722 – 726 (2007).
[16] Sengupta S and Bandyopadhyay S, Evolving fragments to lead molecules in ISB '10: Proceedings of the International Symposium on Biocomputing 2010, New York, 1–7.
[17] Bandyopadhyay S and Sengupta S, "IVGA3D: De novo ligand design using a variable sized tree representation", Protein & Peptides Letters, 17: 1495–1516 (2010).
[18] Vapnik VN, "Estimation of dependencies based on empirical data," Nauka: Moscow, 1979.
[19] V. N. Vapnik, "The nature of statistical learning theory," New York: Springer, 1995.
[20] V. N. Vapnik, Statistical learning theory, Adaptive and learning systems for signal processing, communications, and control, Wiley: New York, 1998.
[21] Ivanciuc O, In Reviews in Computational Chemistry, Wiley-VCH: Weinheim, Germany, 2007, 291—400.
[22] Burbidge R, Trotter M, Buxton B, and Holden S, "Drug design by machine learning: Support vector machines for pharmaceutical data", Computers and Chemistry, 26: 4–15 (2001).
[23] Doniger S, Hofmann T, and Yeh JJ, "Predicting CNS permeability of drug molecules: Comparison of neural network and support vector machine algorithms", Journal of Computational Biology, 9: 849–864 (2002).
[24] Lengauer T, Lemmen C, Rarey M, and Zimmermann M, "Novel technologies for virtual screening", Drug Discovery Today, 9: 27–34 (2004).
[25] Trotter MWB, Buxton BF, and Holden SB, "Support Vector Machines in combinatorial chemistry", Measurement and Control, 34: 235–239 (2001).
[26] Trotter MWB and Holden SB, "Support Vector Machines for ADME property classification", QSAR and Combinatorial Science, 22: 533–548 (2003).
[27] Warmuth MK, Liao J, Ratsch G, Mathieson M, Putta S, and Lemmen C, "Active learning with support vector machines in drug discovery process", J. Chem. Inf. Comput. Sci., 43: 667–673 (2003).
[28] Jorissen RN and Gilson MK, "Virtual screening of molecular databases using a support vector machine", Journal of Chemical Information and Modeling, 45: 549–561 (2005).

[29] Yao XJ, Panaye A, Doucet JP, Zhang RS, Chen HF, Liu MC, Hu ZD, and Fan BT, "Comparative study of QSAR/QSPR correlations using support vector machines, radial basis function neural networks, and multiple linear regression", Journal of Chemical Information and Computer Sciences, 44: 1257–1266 (2004).
[30] Zernov VV, Balakin KV, Ivaschenko AA, Savchuk NP, and Pletnev IV, "Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions", Journal of Chemical Information and Computer Sciences, 43: 2048–2056 (2003).
[31] Byvatov E and Schneider G, "SVM-Based Feature Selection for Characterization of Focused Compound Collections", Journal of Chemical Information and Computer Sciences, 44: 993–999 (2004).
[32] Muller KR, Rtsch G, Sonnenburg S, Mika S, Grimm M, and Heinrich N, "Classifying drug-likeness' with kernel based learning methods", Journal of Chemical Information and Modeling, 45: 249–253 (2005).
[33] Jacob L, Hoffmann B, Stoven V, and Vert JP, "Virtual screening of gpcrs: An in silico chemogenomics approach", BMC Bioinformatics, 9: 363–379 (2008).
[34] Li L, Li J, Khanna M, Jo I, Baird JP, and Meroueh SO, "Docking to erlotinib off-targets leads to inhibitors of lung cancer cell proliferation with suitable in vitro pharmacokinetics", ACS Medicinal Chemistry Letters, 1: 229–233 (2010).
[35] Li L, Khanna M, Jo I, Wang F, Ashpole NM, Hudmon A, and Meroueh SO, "Target-specific support vector machine scoring in structure-based virtual screening: Computational validation, in vitro testing in kinases, and effects on lung cancer cell proliferation", Journal of Chemical Information and Modelling, 51: 755–759 (2011).
[36] Smalter A, Huan J, and Lushington G, "Graph wavelet alignment kernels for drug virtual screening", Journal of Bioinformatics and computational biology, 7: 473—497 (2009).
[37] Soman ST and Soman KP, Wavelet assignment graph kernel for drug virtual screening in Advances in Recent Technologies in Communication and Computing 2009, 282–284