

Applications of Support Vector Machines as a Robust tool in High Throughput Virtual Screening

Renu Vyas, S.S. Tambe, B.D. Kulkarni*

Chemical Engineering and Process Development Division
National Chemical Laboratory, Homi Bhabha Road, Pune - 411008, INDIA

Article Info

Article history:

Received Jan 10th, 2012

Revised Jun 5th, 2012

Accepted Jun 20th, 2012

Keyword:

SVM

Drug designing

Virtual screening

Protein

Statistical methods

Molecular screening

Cancer therapy

ABSTRACT

Chemical space is enormously huge but not all of it is pertinent for the drug designing. Virtual screening methods act as knowledge-based filters to discover the coveted novel lead molecules possessing desired pharmacological properties. Support Vector Machines (SVM) is a reliable virtual screening tool for prioritizing molecules with the required biological activity and minimum toxicity. It has to its credit inherent advantages such as support for noisy data mainly coming from varied high-throughput biological assays, high sensitivity, specificity, prediction accuracy and reduction in false positives. SVM-based classification methods can efficiently discriminate inhibitors from non-inhibitors, actives from inactives, toxic from non-toxic and promiscuous from non-promiscuous molecules. As the principles of drug design are also applicable for agrochemicals, SVM methods are being applied for virtual screening for pesticides too. The current review discusses the basic kernels and models used for binary discrimination and also features used for developing SVM-based scoring functions, which will enhance our understanding of molecular interactions. SVM modeling has also been compared by many researchers with other statistical methods such as *Artificial Neural Networks*, *k-nearest neighbour (kNN)*, *decision trees*, *partial least squares*, etc. Such studies have also been discussed in this review. Moreover, a case study involving the use of SVM method for screening molecules for cancer therapy has been carried out and the preliminary results presented here indicate that the SVM is an excellent classifier for screening the molecules.

Copyright © 2012 *International Journal for Computational Biology*,
<http://www.ijcb.in>, All rights reserved.

Corresponding Author:

B.D. Kulkarni
National Chemical Laboratory,
Homi Bhabha Road,
Pune, India
Email: bd.kulkarni@ncl.res.in



How to Cite:

Renu Vyas *et. al.* Applications of Support Vector Machines as a Robust tool in High Throughput Virtual Screening. IJCB. 2012; Volume 1 (Issue 1): Page 43-55.

1. INTRODUCTION

Virtual screening is a broad term encompassing various tools, techniques and methods to obtain the molecule with desired properties [1]. The vast amount of data being generated by the "omics" technologies and the computational possibility of generating innumerable compounds require a virtual screening (VS) tool capable of identifying different types of potential inhibitors from large compound libraries with high yields and low false-hit rates similar to high throughput screening (HTS). A right choice of the method depends upon the amount of available knowledge at hand. The target-based virtual screening methods such as docking and pharmacophore modeling are routinely used if we have well resolved X-ray crystallographic data available or a good homology model can be generated. In the absence of both, the ligand-based techniques that can be used include Quantitative Structure Activity Relationship (QSAR), pharmacophore searching, similarity searching, database screening etc.[2]. Recently, several machine learning methods have been reported; among them

Support Vector Machine (SVM) and its regression performing analogue namely Support Vector Regression (SVR) have emerged as the most powerful techniques in computational chemistry [3]. Support vector machines have been used on million dimensional data sets and in other cases with more than a million examples [4]. An SVM-based virtual screening method is a rapid computational tool for the prediction of potential lead molecules for any drug discovery program [5]. The SVM and SVR techniques have been extensively employed in bioinformatics and chemical engineering (see e.g. [6-12]). As compared to the SVM, its regression analogue SVR has been sparingly used in virtual screening efforts. Accordingly, this review mainly considers SVM as a powerful technique for speeding up virtual screening efforts. Although there exists a deficiency of 3D structural experimental data on known expressed proteins, the SVM models can be still employed as filters to obtain efficient lead molecules in virtual screening cycles. The SVM builds predictive models for regression and classification of molecules thus offering medicinal chemists a rapid way to identify novel leads requiring least experimental efforts. Its predictive ability is emphasized by the formalism's successful application to different test sets of diverse and various multi-targeted compounds. The SVM models generally give an estimated sensitivity of greater than 83% and specificity of greater than 99%, and thus have been used to screen millions of compounds in free and proprietary chemical databases [13]. They are capable of identifying novel inhibitors and distinguishing inhibitors from structurally similar non-inhibitors. In comparison with other virtual screening tools the SVM model is found to possess a broad applicability domain and a low false positive rate, which makes it suitable for the virtual screening of chemical libraries. SVM models exhibit good accuracy at cross-validation and independent testing. The predictive ability of these models in virtual screening is evaluated by parameters such as the number of false positives, false negatives, true negatives and true positives, sensitivity and specificity, error rate and Mathews correlation coefficient (MCC), and finally experimental validation is conducted to substantiate model predictions. Target-based SVM models can correctly indicate interactions with residues as well as hydrogen bonding between the key residues which has a tremendous relevance for important substructures and functional groups that are linked to the protein-ligand interactions for structure-activity relationship (SAR) studies. Hence, these models consistently provide guidelines for enhancing activity in novel candidate molecules—the ultimate objective of drug designing. In this review, we analyze three major types of SVM-based virtual screening approaches as depicted in Figure 1.

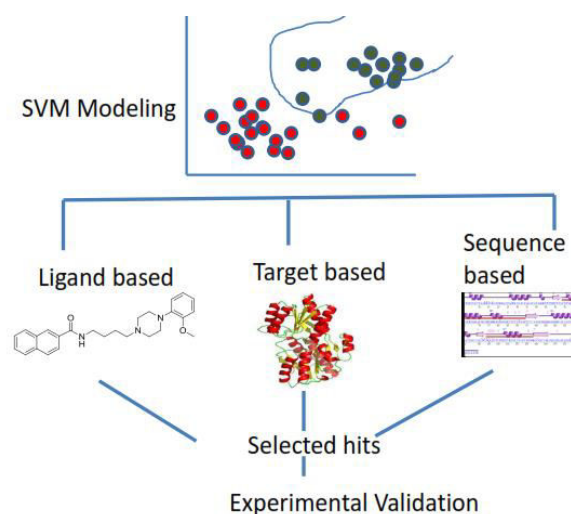


Fig. 1: SVM based major virtual screening methods

1.1 SVM theory: An overview

The SVM is a statistical learning theory based nonlinear model developed by Vapnik et al. [3] that discriminates between data points of distinct classes (binary SVM) such that the margin between both the classes is maximized. This margin models the linear decision hyperplane. The basic idea of SVM is to map the training data nonlinearly into a higher dimensional feature space via a mapping function and construct a separating hyperplane with maximum margin there. This yields a nonlinear decision boundary in the input space.

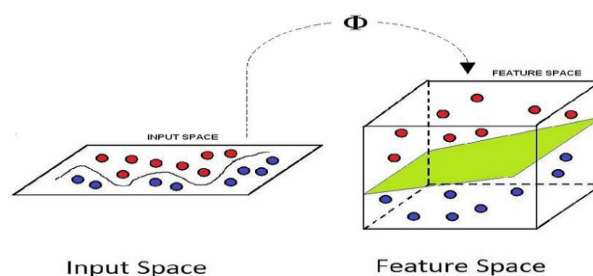


Figure 2: Mapping function Φ maps the training data nonlinearly into a higher dimensional feature space wherein a separating hyperplane with maximum margin is constructed yielding a nonlinear decision boundary in the input space.

The final position and orientation of the hyperplane are defined by a subset of training vectors, the so-called support vectors. Usually, the SVM approaches are used in association with a radial basis function (RBF) as the kernel function although other kernel functions such as dot, polynomial, sigmoid, multi-quadratic, Gaussian combination etc. have also been explored.

SVMs use an implicit mapping, Φ , of the input data into a high-dimensional feature space defined by a kernel function, i.e., a function returning the inner product ($\Phi(x), \Phi(x')$) between the images of two data points x, x' in the feature space. The learning then takes place in the feature space, and the data points only appear inside dot products with other points [14]. This is often referred to as the “kernel trick” [15]. Owing to this trick, it becomes unnecessary to perform any computations in the high-dimensional feature space since all the requisite computations can be carried out directly in the input space.

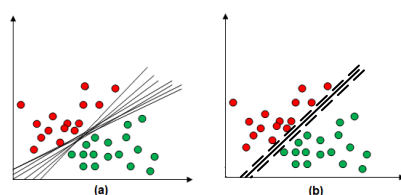


Figure 3 (a) A number of hyperplanes are possible to separate the two classes, (b) The SVM algorithm constructs a hyperplane that maximizes separation between two classes.

As shown in Figure 3 an SVM classifier finds a hyper-plane that maximizes the separation between the two (or more) classes. The panel (a) in Figure 3 shows a number of hyper-planes that separate the data corresponding to a binary classification problem. However, there exists only one hyperplane that maximizes the separation. Locating such a maximum-margin hyperplane (MMH), as shown in the panel (b) of the Figure, results in a classifier possessing improved generalization capability owing to which the classifier can make correct predictions for new inputs. Correctly locating the MMH requires solving a quadratic optimisation problem with a single minimum. As a result, difficulties in locating an optimum solution that corresponds to the global or a deepest local minimum on the error surface (as in Artificial Neural Network training) are avoided in the SVM training. A short summary of the commonly employed SVM training algorithm for correctly placing a large margin hyper-plane is given below [16]:

Given labeled data pairs $(x_1, y_1), \dots, (x_i, y_i)$ comprising multiple input vectors x_i ($i = 1, 2, \dots, n$) and the corresponding scalar outputs, y_i ($i = 1, 2, \dots, n$) and a kernel, k , the SVM computes a function:

$$f_s = \sum_{i=1}^n \alpha_i k(x_i, x) + b$$

where coefficients α_i (Lagrange multipliers) and b are found by solving the following

optimisation problem:

$$\text{Minimize} \quad \sum_{i,j=1}^n \alpha_i \alpha_j k(s_i, s_j) + C \sum_{i=1}^n \xi_i \quad (2)$$

$$\text{subject to} \quad y_i f(x_i) \geq 1 - \xi_i \quad (3)$$

where C refers to the regularization parameter, and ξ_i refers to the slack variables. Intuitively, the kernel function k computes similarity between two given examples. The most commonly employed kernel functions are:

$$k(x, x') = \exp \left(-\frac{\|x - x'\|^2}{\rho^2} \right)$$

RBF kernels: (4)

$$\text{Polynomial kernels:} \quad k(x, x') = (x \bullet x')^d \quad (5)$$

The SVM finds a large margin separation between the training examples and previously unseen examples will often be close to the training examples. The large margin then ensures that these examples are correctly classified as well (high generalization ability). The SVM training algorithm constructs models that are adequately complex yet unlike artificial neural networks are simple enough to be analysed mathematically. SVM is capable of handling multiple continuous and also categorical variables.

1.2 Ligand-based virtual screening using SVM

Since a very few protein crystal structures have been solved experimentally by an X-ray or an NMR technique, the ligand-based approaches serve as a valuable tool for virtual screening. The ligand-based SVM approach can be used even in the absence of receptor information and a number of researchers have used it successfully to obtain inhibitors for a range of pharmacologically important targets. Commonly, two kernels are used in the ligand-based virtual screening viz. 2D Tanimoto kernel and 3D pharmacophore kernel. While a number of graph kernels are used in chemoinformatics, the Tanimoto kernel is the most commonly used one. If we consider feature map ϕ_d and corresponding kernel k_d then Tanimoto kernel k_{td} is defined as [17] k_{td}

$$k_{td}(u, v) = \frac{k_d(u, v)}{k_d(u, u) + k_d(v, v) - k_d(u, v)} \quad (6)$$

where u and v denote two molecules and d is an integer.

The initial reports on virtual screening used Binary Kernel Discrimination (BKD), a fingerprint based method based on Tanimoto concept of similarity as the machine learning approach. Willet et al. [18] performed a comparative study of virtual screening using binary kernel discrimination (BKD) with other ligand fingerprint based virtual screening chemoinformatics methods and found the BKD method to be much superior in drug and pesticide discovery but lower in performance than the SVM. Byvatov et al. [19] trained an SVM for the prediction of D3 and D2 receptor selective ligands. The hit compounds were synthesized and shown to possess nanomolar affinity. This approach was further refined by Chen et al. [20] who observed the effect of noisy data on the model using MDDR database and also noticed that its predictive ability depended on the number of false positives in the training set. Wilton et al. [21] obtained similar results in their work on pesticide data from Syngenta corporate database. Pharmacophores are important structural features present in a molecule which help in binding with a receptor thus giving rise to its biological activity. The pharmacophoric features are limited in number that mainly include ring centroid, aromatic ring, hydrophobic regions, hydrogen bond acceptor, hydrogen bond donors, etc. An example of a three point pharmacophore is shown in Figure 4.

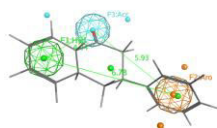


Fig. 4: A three point pharmacophore query and three interatomic distances

Pharmacophores serve as very important parameters in virtual screening of chemical libraries. An SVM based pharmacophore modeling has been performed by using the pharmacophore kernel which is defined as follows [22].

For any pair of pharmacophores, $p = [(x_1 l_1), (x_2 l_2), (x_3 l_3)]$ and $p' = [(x_1' l_1'), (x_2' l_2'), (x_3' l_3')]$

$$K_i(p, p') = \prod_{i=1}^3 K_{Feat}(l_i, l_i') \quad (7)$$

$$K_s(p, p') = \prod_{i=1}^3 K_{Dist}(\|x_i - x_{i+1}\|, \|x_i - x_{i+1}'\|) \quad (8)$$

where $\|\bullet\|$ corresponds to the Euclidean distance, the index $i+1$ is taken as modulo 3, K_{Feat} and K_{Dist} are the kernel functions introduced to compare pairs of labels of atoms and pairs of distances, K_i corresponds to the intrinsic similarity and K_s corresponds to the spatial similarity of the pharmacophores. Franke et al. [23] have demonstrated the effectiveness of this strategy by applying it for screening of COX-2 inhibitors.

Chen et al. [24] used Atom Pair (AP) structure and physicochemical (PC) descriptors of compounds to generate SVM-AP (support Vector Machine - Atom Pairs) and SVM-PC (Support Vector Machine Physico-Chemical) models to develop "LigSeeSVM," a screening tool for ligand-based virtual screening, which was validated on five different datasets. In another study [25], SVM modeling was performed to screen inhibitors for LCK1- a target implicated in the auto-immune diseases. The model had an estimated sensitivity of greater than 83% and specificity of greater than 99%, and it was used to screen 168014 compounds in the MDDR database and found to have a yield of 45.8% and a false positive rate of 0.52 %. Ma et al. [26] assessed the performance of SVM by using a sparse data set of active compounds for six target classes in the MDDR database namely muscarinic M1 receptor agonists, NMDA receptor antagonists, thrombin inhibitors, HIV protease inhibitors, cephalosporins, and renin inhibitors. They found the SVM method to be superior to Tanimoto coefficient based similarity search methods at giving low false hit rates. The predicted compounds were verified by a cellular activity assay proving SVM to be an efficient method. Jorissen and Gilson [27] developed a modified version of SVM to not only classify molecular data but also enrichment of actives by using a novel method for identifying descriptors and cross-validating them as parameters in training the SVM. The results obtained were better than those based on fingerprints such as binary kernel discrimination. Han et al. [28] improved the performance of SVM for virtual screening of huge libraries by including a number of diverse non-actives in the training dataset. The hit rates and enrichment factors were found to increase dramatically for datasets of HIV protease inhibitors, DHFR inhibitors, dopamine antagonists and CNS active agents and were better than those based on the other ligand-based virtual screening methods.

1.3 Sequence based virtual screening using SVM

In a recent report by Wang et al. [29], a novel SVM method using only sequence data of targets and information on 2D structures of small molecules was developed. This SVM model was based on 15,000 ligand protein interactions derived from 626 protein and 10,000 active compounds. The methodology was used for identifying nine active compounds for four targets viz., GRP40, SIRT1, p38, and GSK-3 β and can also be extended to other proteins.

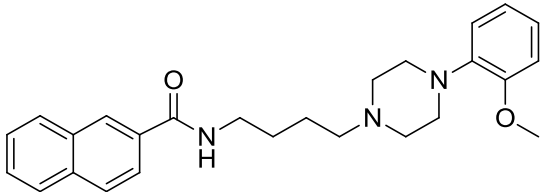
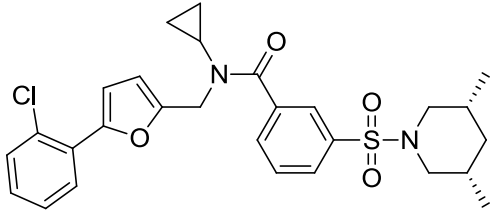
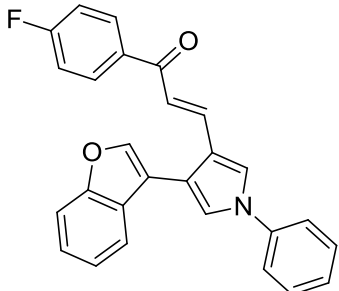
1.4 Target-based virtual screening using SVM

Target-based virtual screening methods have recently been introduced compared to the conventional ligand based methods. The latter method has several drawbacks, the most important being the neglect of the receptor with no consideration to its flexibility and secondly a lack of adequate scoring functions. Currently, a number of research groups are focussing on the target-based SVM methods for success in virtual screening. Li et al. [30] have reported SVM-SP, an exclusive target based scoring function that yielded better enrichment than Glide docking score as evidenced by the ROC-AUC (Receiver Operating Curve - Area Under the Curve) characteristic plots. They evaluated its performance on 41 targets mostly from Directory of Useful Decoys (DUD) and obtained best results with kinases. The strategy worked well with the homology model and also succeeded when few structures were available as the training set. Virtual screening was performed against 1125 compounds for two targets namely EGFR and CAMKII wherein three out of the 25 hit compounds showed good

inhibitory activity in vitro. In another report by the above group [31] the support vector regression method was applied for rank ordering and virtual screening of chemical libraries using community structure-activity resource (CSAR) datasets. Here, two new scoring functions were developed based on the knowledge of pair-wise potential and physiochemical properties. These scoring functions outperformed the well-established seven scoring functions namely Glide, VINA, Gold score, Dock, Chemscore, PMF and X score. Li et al.[31] developed a new function SVR_KBD employing a target specific strategy and the enrichment results were found comparable to their previously reported SVM-SP scoring function.

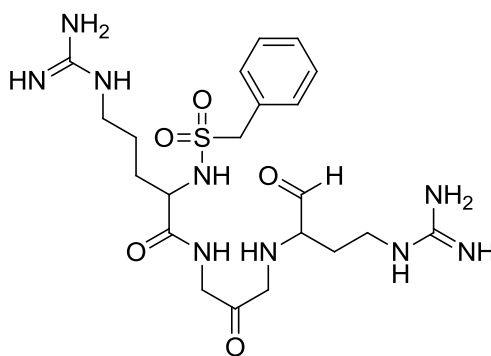
Waaserman and Bajorath [32] have discussed at length various SVM-based target selective searching strategies for virtual screening. They have elucidated a superior approach in terms of enrichment factor as the SVM is trained on the data comprising more than two different classes viz., selective, promiscuously active and non-active compared to the commonly employed binary classification approach. Further, they present a modified preference ranking strategy leading to higher recall of selective compounds. Combinatorial support vector machines have been used as virtual screening tools [33] for searching dual-inhibitors of 11 combinations of 9 anticancer kinase targets (EGFR, VEGFR, PDGFR, Src, FGFR, Lck, CDK1, CDK2, GSK3). In this study, the C-SVM was found either comparable or slightly better than the other conventional method such as Surflex, Dock, Blaster, KNN and PNN. Plewczynski et al. [34] performed an exclusive target-specific supervised SVM analysis for compounds retrieved from MDDR database related to five targets including cyclooxygenase-2, dihydrofolate reductase, thrombin, HIV-reverse transcriptase and antagonists of the estrogen receptor. The SVM model was based on only two dimensional topological descriptors related to atom pairs. The sensitivity and classification for all the protein targets were 80% and 100%, respectively. The literature is replete with more examples of successful application of SVM with both ligand and target based approaches. Table 1 shows a few representative ligands and their corresponding targets; needless to say that SVM modeling has comprehensively mapped chemical diversity and target space in virtual screening.

Table 1: Representative examples of targets and ligands covered by SVM obtained from recent literature

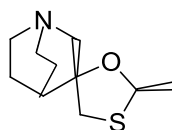
Sl No	Targets	2D structure of selected molecules	Reference
1	D3 dopamine receptor	BP987(a partial agonist) 	Byvatov et al., [19]
2	γ secretase		Xue-Gang et al., [43]
3	EGFR and CamKII		Li et al., [30]

4 Factor Xa

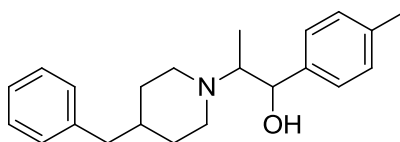
Li et al., [31]



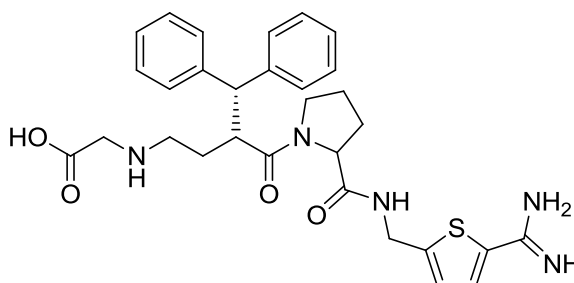
5 Muscarinic M1

Ma X H et al.,
[26]

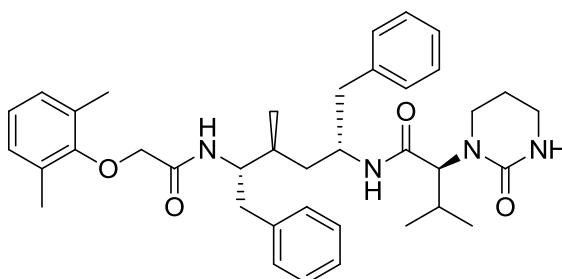
6 NMDA

Ma X H et al.,
[26]

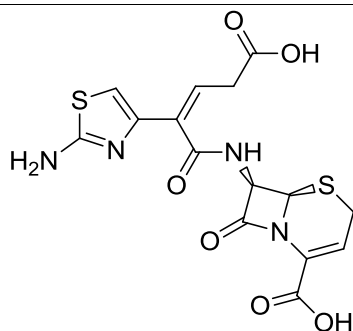
7 Thrombin

Ma X H et al.,
[26]

8 HIV protease

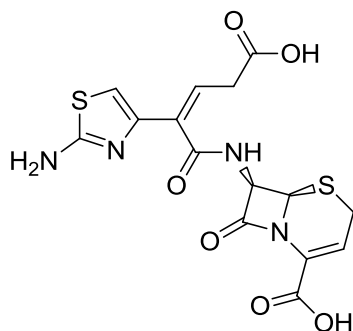
Ma X H et al.,
[26]

9 Cephalosporin

Ma X H et al.,
[26]

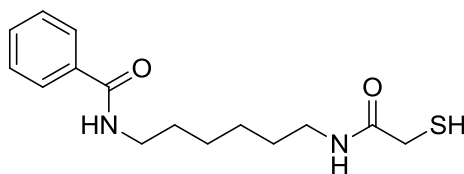
10 Renin

Ma X H et al., [26]



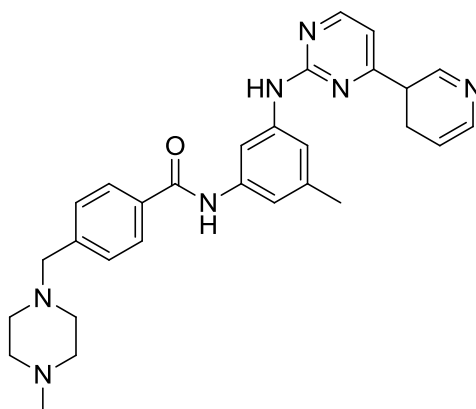
11 HDAC

Tropsha , [45]

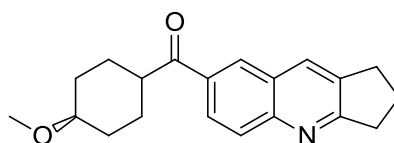


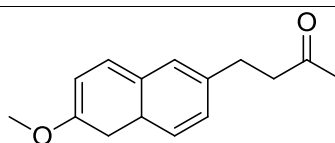
12 AbL

Liu et al., [44]



13 mGluR1

Mballo and
Makarenkov [42]



1.5 Feature based library selection

Although machine learning methods including the SVM yield good classification/regression results they are considered as "black-box" since the form and parameters of the developed model cannot be interpreted meaningfully to gain an insight into the classification/ regression process. In the context of virtual screening, for example, the influence of individual molecular feature on the classification performance is not easily discernable. An interesting attempt to circumvent this drawback was made by Byvatov and Schneider [35] who reported an SVM based algorithm for feature-based selection method for libraries of enzyme inhibitors towards a better understanding of the protein-ligand interactions. The study included a comparison with the classical model-independent Kolmogorov-Smirnov (KS)-based feature selection and it was found that the SVM is faster than the KS-based classifiers. This algorithm established the SVM as more intelligent, comprehensible and no longer a complete black box.

1.6 Comparison of SVM with other virtual screening methods

A number of researchers have used SVM along with other commonly utilized machine learning methods in the arena of virtual screening and the comparative results of some of these studies are briefly discussed here. An SVM in combination with docking studies produced better results for virtual screening experiments as observed by Li and co-workers [36] who identified a novel series of benzimidazole derivatives for EGFR, VEGFR and PDGFR kinases. Similar results have been reported by Xie et al. [37] who discovered new and potent inhibitors of c-Met, a membrane receptor required for embryonic development and wound healing. The SVM results when supplemented with docking results led to higher number of hits and enrichment of actives in the dataset. Some of the SVM predicted actives were also validated experimentally in assays. Ren et al. [38] used a hierarchical methodology encompassing three virtual screening techniques viz. SVM, pharmacophore modeling and docking in succession to predict potent inhibitors for Pim-1. These techniques were used to screen a large number of databases including Pubchem, Specs and Enamine and finally fifteen hits showing nanomolar activity were obtained. Luan and co-workers [39] used this strategy of combining SVM and molecular docking for discovering 9-amino acridine scaffolds as multi-target inhibitors for VEGFR-2 and Src Kinase and confirmed the results experimentally.

Apart from biological activity, toxicity of drugs is another major concern for pharma companies and relatively few accurate prediction methods are available. Kumar et al. [40] used SVM methods to correctly estimate the genotoxicity levels of compounds in their training set. They studied in depth the effect of training set size and noise levels on the performance of SVM analyzing genotoxic and nongenotoxic compounds from large virtual screening libraries. The predictions made by the SVM model were on par with those made by TOPKAT- a commercial toxicity prediction tool by Accelrys. In another study [41], SVM in combination with pharmacophore modeling yielded good results for the development of glutamate non-competitive antagonists of mGluR1, a target implicated in nervous disorders. As the X-ray-solved structure is not available, ligand based virtual screening is the only approach for this target. Using MDDR data set it was shown that multi-step virtual screening approach involving both virtual screening techniques is superior to using each of them individually. Mballo and Makarenkov [42] made a comparative study of six known machine learning methods viz. binary decision trees, neural networks, SVM, linear discriminant analysis, k-nearest neighbours and partial least squares by analysing test assay from the McMaster University Data Mining and Docking Competition. They evaluated the methods on the basis of various parameters such as sensitivity, enrichment factor and number of false positive and negatives. Finally, they came up with a variable selection procedure and applied it to the polynomial SVM. Yang and co-workers [43] used random forest (RF) and SVM learning techniques to design inhibitors for gamma secretase, an important target for Alzheimer's disease. They observed that the RF model marginally outperformed the SVM method. Virtual screening using the model resulted in three hits in the ZINC database. AB1 is an important target for cancer therapy and considerable efforts have been made for developing inhibitors using insilico methods such as docking and pharmacophore modeling. Liu et al. [44] found SVM

approach better than the above mentioned techniques for identifying ABl inhibitors as it led to lower false hit rates and enabled searching of huge libraries. Tropsha et al. [45] used a combined Quantitative Structure Activity Relationship -Virtual Screening (QSAR-VS) approach involving kNN and SVM to develop human histone deacetylase HDAC inhibitors. Highly predictive models with good r^2 values were obtained and were rigorously cross-validated on external datasets. The model gave forty five unique hits while searching a huge in-house database. In another comparative study by Bajorath and co-workers [46] the ranking provided by involving SVM proved to be far superior than the ones provided by nearest neighbour and centroid similarity search methods even when a smaller data set was used for training. The explanation given was that during the learning phase SVM uses information about database molecules, in addition to known active compounds. In a systematic study, Byvatov and co-workers [47] compared ANN and SVM methods as binary classifiers for discriminating drugs from non-drugs. They used three sets of descriptors viz. 120 Ghose-Crippen fragments, a wide range of 180 descriptors from the Molecular Operating Environment (MOE) package, and 225 topological pharmacophore (CATS) descriptors. In general, SVM performed marginally better than the ANN with minimum error regardless of the choice of descriptors. However, the authors concluded that the two methodologies are complementary to each other as the results were similar but not identical. Melagraki et al. [48] did extensive work on developing inhibitors for MCH1 receptor by using a number of ligand based virtual screening techniques in tandem. First, a linear QSAR model was developed using multiple regression method following which the most suitable input variables were selected using the Elimination Selection-Stepwise Regression (ES-SWR) method. Finally, SVM was used to categorize the molecules into actives and non-actives. A number of efforts were expended to select the optimum scaffold and the activities of the predicted actives by SVM were estimated by using the MLRS model. Jorgenson group at University of Copenhagen developed many in silico models based on different classification methods such as binary QSAR, kNN, SVM, decision tree etc. for developing inhibitors for P450 1A2, an important enzyme in drug metabolism. Here, SVM, kNN and random forest methods were found to be the best methods delivering models with high prediction accuracy with a Mathews correlation coefficient of 0.5 [49]. Similar results were obtained by Khandelwal et al. [50] in their work on predicting pregnane X receptor activators using machine learning methods coupled with docking protocol. They observed that docking combined with regression yielded inferior results when compared with SVM and RF methods. Plewczynski et al. [51] have conducted extensive studies to assess a host of machine learning techniques such as SVM, random forest, ANN, k-nearest neighbour (kNN) classification with genetic-algorithm-optimized feature selection, trend vectors, naive Bayesian classification, and decision tree, for their capacity to recognize ligands from a large data collection of molecules. Interestingly, they obtained varying results from the stated methods; while some were good in retrieving actives, others yielded high enrichment scores. However, all the methods could not correctly predict the recently reported ligands. It was concluded that no single method can be the most consistent one; rather a combination of methods is essential for better results in virtual screening.

A case study: SVM based binary classification of molecules for their potential as anticancer agents

A number of molecules tested as anticancer lead compounds against the human breast cancer cell line MCF-7 were downloaded from Pubchem bioassay and NCI database and used for building an SVM based classifier. A dataset of 54 molecules containing actives and in-actives was constructed and split into training set (41 molecules) and test set (13 molecules) (see Figure 5).

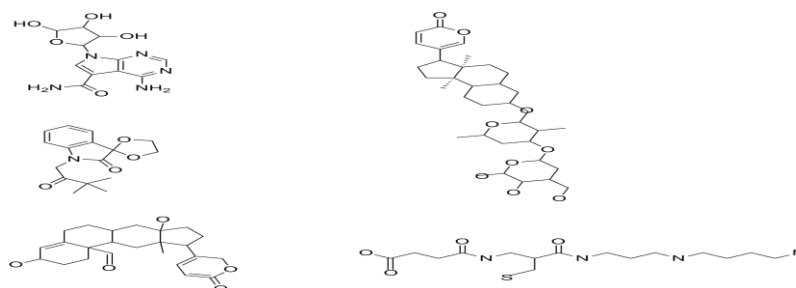


Fig.5: 2D structure of some diverse molecules used in training and test set

The objective of the study was to build a classifier using SVM to discriminate inactives from actives for the given cancer cell line data. The molecules were washed, energy minimized and 186 descriptors were computed using MOE [52]. They include topological, shape and connectivity based descriptors. The descriptors were pruned using the QSAR contingency module in MOE and 20 descriptors mainly the Lipinski RO5 and

BCUT descriptors closely encoding the activity were selected. RapidMiner 5.1 an open source data-mining software tool was used to generate the SVM classifier model [53]. The objective of the classification was to categorize a given molecule into “active” or “inactive” depending upon the descriptors encoding the activity. This is a supervised classification task where an example data containing each molecule’s descriptors (model inputs) and its class (active or inactive) defining model output are available. The RapidMiner consists of various types of SVM modules for performing supervised classification (clustering). Among these, the Support Vector Machine -Evolutionary (SVME) [54] yielded best classification results. The SVME uses an evolutionary algorithm (ES) for solving the dual optimization problem of an SVM. On many datasets, SVME performs as fast and accurate as the usual SVM implementations. Additionally, it is also capable of learning with Kernels which are not positive semi-definite.

The parameter values used in conducting the above stated classification are: (i) kernel type: ANOVA, kernel gamma = 1.0, kernel degree = 3, C = 0.25, ϵ = 0.1, maximum number of generations = 10,000, population size = 5, selection type: Tournament, and crossover probability = 1.0. The stated combination of SVME parameters yielded excellent classification results wherein all the molecules in both the training and test sets were classified with 100% accuracy (no false positives or false negatives). For a training set of 41 input-output patterns, the SVME method identified 24 support vectors. An excellent classification accuracy of 100% in respect of the test set molecules indicates that the developed SVME model is capable of accurately generalizing the learned classification to new molecules. A detailed study comparing the classification results from a number of SVM and ANN based classifiers for other cell lines data is currently in progress.

2. CONCLUSION

SVM based screening is flexible, fast and it significantly increases the speed and accuracy of prediction in virtual screening experiments. To some extent, these advantages are offset by limitations such as low hit rate and high number of false positives. Consequently, there exists a need to speed up SVM and kernel methods, which will surely benefit virtual screening efforts at large. As rightly pointed out by Schneider [55] it might be wise to try out several predictive methods in parallel for the right solution in virtual screening. To sum it up, a consensus approach would be ideal for the large disparate datasets generally available for computational biology studies.

ACKNOWLEDGEMENTS

Renu Vyas thanks Department of Science & Technology, Govt of India, New Delhi, for the award of “Women Scientist Fellowship

REFERENCES

- [1] Walters W.P., Stahl M.T. and Murcko M.A. "Virtual screening – an overview", *Drug Discov Today*, 3: 160–178 (1998).
- [2] Alexander Tropsha and Alexandre Varnek Eds *Cheminformatics approaches to virtual screening* Royal Society of Chemistry 2008.
- [3] Vapnik V., Golowich S. and Smola A., "Support vector method for function approximation, regression estimation and signal processing", *Adv. Neural Inform. Process. Syst.* 9: 281–287 (1996) ; Schölkopf B., Simard, P.Y., Smola, A. J. and Vapnik V. Prior knowledge in support vector kernels In *Advances in Neural Information Processing Systems*; Jordon M., Kearns M. and Solla S. Eds.; MIT Press :Cambridge, MA, 1998; pp 640–646.
- [4] Mangasarian O. L. and Musicant D. R. "Lagrangian Support Vector Machines", *Journal of Machine Learning Research*, 1: 161-177 (2001).
- [5] Ovidiu Ivanciuc Ed Lipkowitz and T.R. Cundari Application of Support Vector Machines in Chemistry In *Reviews in Computational Chemistry*; Wiley-VCH, Weinheim 2007 ; pp 291-40.
- [6] Kulkarni A., Jayraman V.K. and Kulkarni B.D "Control of chaotic dynamical systems using support vector machines", *Physics Letters A*, 317: 429-435 (2003).
- [7] Mundra P.K., Kumar M.K., Krishna K.K., Jayraman V.K. and Kulkarni B.D. "Using psuedoamino acid composition to predict protein subnuclear localization: Approached with PSSM", *Pattern Recognition Letters* 28: 1610-1615 (2007).
- [8] Kulkarni A., Jayraman V. K. and Kulkarni B.D. "Support vector classification with parameter tuning assisted by agent based technique", *Computers and Chemical Engineering* 28: 311-318 (2004).
- [9] Kulkarni A., Jayraman V.K. and Kulkarni B.D. (2005) "Knowledge incorporated support vector machines to detect faults in Tennessee Eastman Process", *Computers and Chemical Engineering*, 29: 2128-2133(2005).
- [10] Jade A.M., Srikanth B., Jayraman V.K., Kulkarni B.D. and Jog J.P "Feature extraction and denoising using kernel PCA", *Chemical Engineering Sciences* 58:4441-4448 (2003).

- [11] Nandi S., Badhe Y., Lonari J., Sridevi U., Rao B. S., Tambe S.S. and Kulkarni, B.D. "Hybrid Process Modeling and optimization strategies integrating neural network/support vector regression and genetic algorithms: Study of benzene isopropylation on H beta catalyst", *Chemical Engineering Journal* 97: 115-129 (2004).
- [12] Gandhi A. B., Joahi J.B., Jayraman V.K. and Kulkarni B.D. "Development of support vector regression(SVR) based correlation for prediction of overall gas hold up in bubble column reactors for various gas-liquid system", *Chemical Engineering Science*, 62: 7078-7089 (2007).
- [13] Schierz A.C. "Virtual screening of bioassay data", *Journal of Chemoinformatics*, 1:21 (2009) .
- [14] Karatzoglou A., Meyer D. and Hornik K. "Support Vector Machines in R", *Journal of Statistical Software*, 15: 1-28 (2006).
- [15] Scholkopf B. and Smola A. *Learning with Kernels*. MIT Press Cambridge MA, 2002.
- [16] Rätsch G. A brief introduction into machine learning In 21st Chaos Communication Congress, Berliner Congress Center, Berlin, Germany 2004 .
- [17] Ralaivola L., Swamidass S. J., Saigo H. and Baldi P. "Graph Kernels for Chemical Informatics", *Neural Networks*, 18: 1093-110 (2005).
- [18] Willett P., Wilton D., Basil H., Tang R., Ford J. and Madge D. "Prediction of Ion Channel Activity Using Binary Kernel Discrimination" *Journal of Chemical Information and Modeling*, 47: 1961-1966 (2007).
- [19] Byvatov E., Sasse B.C., Stark H. and Schneider G. "From virtual to real screening for D3 dopamine receptor ligands", *ChemBioChem*, 6 : 997-999 (2005).
- [20] Chen B., Harrison R.F., Hert J., Mpanhanga C., Willett P. and Wilton D. J. "Ligand-based virtual screening using binary kernel discrimination", *Molecular Simulation*, 31: 597-604 (2005).
- [21] Wilton D.J., Harrison R.F., Willett P., Delaney J., Lawson K., and Mullier G "Virtual Screening Using Binary Kernel Discrimination: Analysis of Pesticide Data", *Journal of Chemical Information and Modeling*, 46: 471-477 (2006).
- [22] Mahe P., Ralaivola L., Stoven V. and Vert J.P. "The pharmacophore Kernel for Virtual Screening with Support Vector Machines", *Journal of Chemical Information and Modeling*, 46: 2003-2014 (2006).
- [23] Franke L., Byvatov E., Werz O., Steinhilber D., Schneider P. and Schneider G. "Extraction and visualization of potential pharmacophore points using support vector machines: application to ligand-based virtual screening for COX-2 inhibitors", *J Med Chem*, 48: 6997-7004 (2005).
- [24] Chen Y.F., Hsu K.C., Lin P.T., Hsu D.F., Kristal B.S. and Yang J.M. "LigSeeSVM: ligand-based virtual screening using support vector machines and data fusion", *Int J Comput Biol Drug Des*, 4: 274-89 (2011).
- [25] Liew C.Y., Ma X.H., Liu X., Yap C.W. "SVM model for virtual screening of Lck inhibitors", *J Chem Inf Model*. 49: 877-85 (2009).
- [26] Ma X.H., Wang R., Yang S.Y., Li Z.R. , Xue Y., Wei Y.C., Low B.C. and Chen Y.Z. "Evaluation of virtual screening performance of support vector machines trained by sparsely distributed active compounds ", *J Chem Inf Model* 48:1227-37 (2008).
- [27] Jorissen R.N. and Gilson M.K. "Virtual screening of molecular databases using a support vector machine", *J Chem Inf Model*, 45:549-61 (2005).
- [28] Han L.Y., Ma X.H., Lin H.H., Jia J., Zhu F., Xue Y., Li Z.R., Cao Z.W., Ji Z.L. and Chen Y.Z. "A support vector machines approach for virtual screening of active compounds of single and multiple mechanisms from large libraries at an improved hit-rate and enrichment factor", *J Mol Graph Model*. 26: 1276-86 (2008).
- [29] Wang F., Liu D., Wang H., Luo C., Zheng M., Liu H., Zhu W., Luo X., Zhang J. and Jiang. "Computational Screening for Active Compounds Targeting Protein Sequences: Methodology and Experimental Validation", *J Chem Inf Model*, 51: 2821-2828 (2011).
- [30] Li L., Wang B. and Meroueh S.O. "Support vector regression scoring of receptor-ligand complexes for rank-ordering and virtual screening of chemical libraries", *J Chem Inf Model*. 51: 2132-8 (2011).
- [31] Li L., Khanna M., Jo I., Wang F., Ashpole N.M., Hudmon A. and Meroueh S.O. "Target specific support vector machine scoring in structure-based virtual screening : computational validation, invitro testing in kinases and effects on lung cancer cell proliferation", *J Chem Inf Model* 51, 755-759 (2011).
- [32] Wassermann A.M., Geppert H. and Bajorath J. "Application of support vector machine-based ranking strategies to search for target-selective compounds", *Methods Mol Biol*. 672, 517-30 (2011) .
- [33] Ma X.H., Wang R., Tan C.Y., Jiang Y.Y., Lu T., Rao H.B., Li X.Y., Go M.L., Low B.C. and Chen YZ "Virtual Screening of Selective Multitarget Kinase Inhibitors by Combinatorial Support Vector Machines", *Mol Pharm.*,7: 1545-1560 (2010).
- [34] Plewczynski D., von Grothuss M., Spieser S.A., Rychlewski L., Wyrwicz L.S., Ginalski K. and Koch U. "Target specific compound identification using a support vector machine" , *Comb Chem High Throughput Screen*, 10: 189-96 (2007).
- [35] Byvatov E. and Schneider G. "SVM-based feature selection for characterization of focused compound collections", *J Chem Inf Comput Sci*, 44: 993-9 (2004).
- [36] Li Y., Tan C., Gao C., Zhang C., Luan X., Chen X., Liu H., Chen Y and Jiang Y. "Discovery of benzimidazole derivatives as novel multi-target EGFR, VEGFR-2 and PDGFR kinase inhibitors", *Bioorg Med Chem*, 19: 4529-35 (2011) .
- [37] Xie Q.Q., Zhong L., Pan Y.L., Wang X.Y., Zhou J.P., Di-Wu L., Huang Q., Wang Y.L., Yang L.L., Xie H.Z. and Yang S.Y. "Combined SVM-based and docking-based virtual screening for retrieving novel inhibitors of c-Met", *Eur J Med Chem*, 46: 3675-80 (2011).
- [38] Ren J.X., Li L.L., Zheng R.L., Xie H.Z., Cao Z.X., Feng S., Pan Y.L., Chen X., Wei Y.Q. and Yang S.Y. "Discovery of novel Pim-1 kinase inhibitors by a hierarchical multistage virtual screening approach based on SVM model, pharmacophore, and molecular docking", *J Chem Inf Model*, 1: 1364-75 (2011) .

- [39] Luan X., Gao C., Zhang N., Chen Y., Sun Q., Tan C., Liu H., Jin Y. and Jiang Y. "Exploration of acridine scaffold as a potentially interesting scaffold for discovering novel multi-target VEGFR-2 and Src kinase inhibitors", *Bioorg Med Chem*, 19: 3312-9 (2011).
- [40] Kumar P., Ma X., Liu X., Jia J., Bucong H., Xue Y., Li Z.R., Yang S.Y., Wei Y.Q. and Chen Y.Z.. "Effect of training data size and noise level on support vector machines virtual screening of genotoxic compounds from large compound libraries", *J Comput Aided Mol Des*, 25: 455-67 (2011).
- [41] Guo-Bo L., Ling L. Y., Shan F., Jian P. Z. , Huang Q, Zhang H.X., Lin Li. Li. and Sheng Y.Y. "Discovery of novel mGluR1 antagonists : A multistep virtual screening approach based on a SVM model and a pharmacophore hypothesis significantly increases the hit rate and enrichment factor", *Bioorganic and Medicinal Chemistry Letters* 21:1736-1740 (2011).
- [42] Mballo C. and Makarenkov V. "Using machine learning methods to predict experimental high-throughput screening data", *Comb Chem High Throughput Screen*, 13: 430-41 (2010).
- [43] Yang G.X., Wei L.V., Yu-Z. C. and Ying X. "Insilico prediction and screening of γ secretase inhibitors by molecular descriptors and machine learning methods", *J Comput Chem*, 31: 1249-58 (2010).
- [44] Liu X.H., Ma X.H., Tan C.Y., Jiang Y.Y., Go M.L., Low B.C. and Chen Y.Z. "Virtual screening of Abl inhibitors from large compound libraries by support vector machines", *J Chem Inf Model*, 49: 2101-10 (2009).
- [45] Tang H., Wang X.S., Huang X.P., Roth B.L., Butler K.V., Kozikowski A.P, Jung M. and Tropsha A. "Novel inhibitors of human histone deacetylase (HDAC) identified by QSAR modeling of known inhibitors, virtual screening, and experimental validation", *J Chem Inf Model* , 49: 461-76 (2009).
- [46] Geppert H., Horváth T., Gärtner T., Wrobel S. and Bajorath J. J. " Support-vector-machine-based ranking significantly improves the effectiveness of similarity searching using 2D fingerprints and multiple reference compounds", *Chem Inf Model*, 48: 742-6 (2008).
- [47] Byvatov E., Fechner U., Sadowski J., Schneider G. "Comparison of support vector machine and artificial neural network systems for drug/nondrug classification", *J Chem Inf Comput Sci* , 43: 1882-9 (2003).
- [48] Melagraki G., Afantitis A., Sarimveis H., Koutentis P.A., Markopoulos J. and Igglessi-Markopoulou O. "Optimization of biaryl piperidine and 4-amino-2-biarylurea MCH1 receptor antagonists using QSAR modeling, classification techniques and virtual screening", *J. Comput Aided Mol Des*, 21: 251-67 (2007).
- [49] Vasanathanathan P., Taboureaux O., Oostenbrink C., Nico P. E., Olsen V.L. and Jørgensen F.S. "Classification of cytochrome P450 1A2 inhibitors and non inhibitors by machine learning techniques", *Drug Metabolism and Disposition* 37 : 658-664 (2009).
- [50] Khandelwal A., Krasowski M.D., Reschly E.J., Sinz M.W., Swaan P.W. and Ekins S. "Machine Learning Methods and Docking for predicting Human Pregnane X receptor activation", *Chem Res Toxicol*, 21: 1457-1467 (2008) .
- [51] Plewczynski D., Spieser, S.A.H. and Koch U. "Assessing Different Classification Methods for Virtual Screening", 46: 1098-1106 (2006).
- [52] MOE www.chemcomp.com/software.htm
- [53] Mierswa I. and Wurst, Michael and Klinckenberg, Ralf and Scholz, Martin and Euler, Timm: "YALE: Rapid Prototyping for Complex Data Mining Tasks", In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-06), 2006.
- [54] Back T., Fogel D. and Michalewicz Z. Eds Handbook of evolutionary computation, Institute of Physics publishing and oxford university press, New York, 2007.
- [55] Schneider G. " Virtual screening: an endless staircase?", *Nature Reviews Drug Discovery*, 9: 273-276 (2010).