NCACMP 18
THE EFFECTIVENESS OF NORMALIZED MOLECULAR DESCRIPTORS IN PREDICTING THE DRUGLIKENESS OF MOLECULES

B. Jeevarathinam[a], T.V.Sundar[*b], G. Vinotha[c], R. Subbiramanian[a]
[a]Postgraduate and Research Dept. of Physics, National College (Autonomous), Tiruchirappalli-620001, India.
[b]Associate Professor, PG & Research Dept. of Physics, National College (Autonomous), Tiruchirappalli-620001, India.
[c]Research Scholar, Postgraduate and Research Dept. of Physics, National College (Autonomous), Tiruchirappalli-620001, India.

Abstract: ***Among the various routines adopted in the drug development process, computer based screening of potential candidate drug molecules is an important processing stage. The goal of screening is to identify suitable lead molecules and to subject them into the further stages of drug development. The screening could be made fast and reliable by making use of numerical data derived from the candidate molecules and by employing machine learning algorithms like support vector machine. The aim of the work is to map the structure based molecular descriptors in to a high dimensional feature space and to subsequently carry out a linear regression in the feature space. Normalized data sets, in the scaled range between zero and one, are quite effective in the pre-processing stages of machine learning algorithms. Hence, in this work such normalized molecular structure descriptors are used to standardize the independent descriptor data sets of candidate molecules. The molecules belong to anticonvulsant category and the efficiency of anova, polynomial and radial kernels are studied. The results of the prediction are compared with the predictions made by the use of directly derived descriptors, normalized ones and with the predictions made by drug mint, a webserver tool.***

Key words: Virtual screening; SVM; anova; polynomial; radial kernels.

---

1. Introduction

Virtual screening (VS) is a computational technique used in drug discovery process. By using computers, a quick search of large libraries of chemical structures can be tested in order to select the structures which could easily bind to a drug target. For the rapid VS process of molecules, a suitable set of Molecular Descriptors (MDs) can be used along with data mining algorithms. The aim of virtual screening is to identify molecules of novel chemical structure or utility in the drug discovery/ development process so that the entire process time would reduce appreciably. According to Todeschini & Consonni [1], "The molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment". A MD is simply a numerical value that could be derived from molecular information sources like chemical formula, molecular structure, its interaction with other molecules etc. The MDs play a fundamental role in chemistry, pharmaceutical sciences, environmental protection policy, health researches and quality control. Different types of descriptors like constitutional, topological, electronic and geometric (2-D or 3-D) could be computed for the molecules by making use of the respective properties of them. They can be used to generate either Quantitative Structure Property Relationships (QSPR) or Quantitative Structure Activity Relationships (QSAR) so that the physico-chemical properties and biological activities of the molecules can be theoretically analyzed. Both QSPR and QSAR attempt to correlate structural or property descriptors of compounds with activities. The main assumption in QSAR is that structurally similar molecules tend to have similar activities and that molecules with unknown properties can be compared to structures with known properties. The work flow involves numerical characterization of molecules on the basis of a particular type of descriptors and subsequent computational experiments. The process is illustrated in Figure 1. The derived mathematical expression or numerical results, if carefully validated can then be used to predict the modeled response of other chemical structures, by carefully verifying them in the applicability domain. Precisely, Molecular Activity = function (physicochemical and /or structural properties) ± Error.

2. Materials and Methods

2.1. Basics of SVM

In the light of the above mentioned background, we have designed and tested a screening strategy by making use of the numerical data derived from the geometry of the molecular structures and with a data mining algorithm called Support Vector Machine (SVM). The SVM is a state-of-the-art classification method introduced in 1992 by Boser, Guyon, and Vapnik [2].

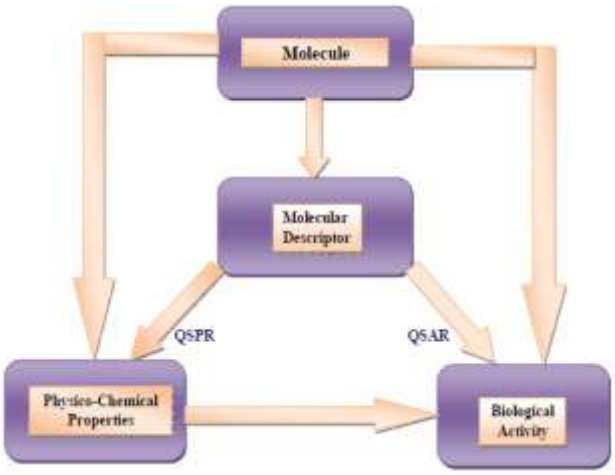

Figure 1: Routes of Virtual Screening with Molecular Descriptors

The SVM classifier is widely used in bioinformatics and other disciplines due to its high accuracy, ability to deal with high-dimensional data such as gene expression, and for its flexibility in modelling diverse sources of data. SVMs belong to the general category of kernel methods [3]. A kernel method is an algorithm that depends on the data only through dot-products. When this is the case, the dot product can be replaced by a kernel function which computes a dot product in some possibly high dimensional feature space [4]. This has two advantages: First, the ability to generate non-linear decision boundaries using methods designed for linear classifier. Second, the use of kernel functions allows the user to apply a classifier to data sets that have no obvious fixed dimensional vector space representation. Linear kernels have limited use but non-linear kernels can be tested in diverse classification problems. Some of the frequently used non-linear kernels are

Polynomial: $\mathbf{K(u,v) = (u.v + 1)^p}$

Gaussian Radial Basis Function: $\mathbf{K(u,v) = \exp(-\| u\text{-}v \|^2 / 2\sigma^2)}$

Sigmoidal (hyperbolic separating surface): $\mathbf{K(u,v) = \tanh(ku\ .v\text{-}\delta)}$

Anova: $\mathbf{K(x,y) = {}_{i=1}\Pi^n\, k(x^i, y^i)}$

The aim of the method is to map the structural descriptors (input data) into a high dimension feature space and subsequently carry out the linear regression in the feature space using suitable kernel functions.

2.2. Experiment

2.2.1. Feature selection and test data

The feature selection of the experimented molecules are based on the application of the data set to figure out the set of physico-chemical properties related to molecular structure, calculated by swissADME [5] based on the calculation 10 sets of structure based descriptors are used to develop the SVM models. For this, it is necessary to use the optimization parameters **(C, γ)** for the kernel functions and these values for each set of descriptors were found out based on MSE values, obtained after the optimization of the training set data. The structure based descriptors used in the study are No. of heavy atoms, No. of aromatic heavy atoms, No. of rotatable bonds, No. of hydrogen bond acceptors, No. of hydrogen bond donors, Molar refractivity and Total Polar Surface Area (TPSA in $Å^2$), Molecular weight, GI absorption and Consensus Log P. A total of 77 molecules for the experiment are taken from Drug Bank [6] database and belonged to approved/ investigational drugs of three categories. The details are given in Table 1.

2.2.2. Classification of data sets

The 77 molecules chosen from Drug Bank repository belonging to the drug category of anticonvulsants are subjected to structure based descriptor calculations and druglikeness filter tests (Lipinski, Ghose, Veber, Egan and Muegge filters) using swissADME. The calculations gave 770 descriptors for analysis. For the SVM training phase requirement they are classified into two classes. Those molecules which passed all the fire filter test are put under first category with class +1 (Drug like). The rest of the molecules even if failed in one filter test or more are put under the second category with class -1(non-Drug like). This classification is done in order to facilitate separation by the hyper plane during machine learning.

2.2.3. Preparation of data sets for modelling experiments

The SVM modelling involves three phases viz, training, validation and testing. First a set of sixteen molecules have been selected randomly from the data set and are considered as the external test set and accommodated both class +1 and class -1 molecules, twelve for +1 category and four for -1 category. This has been done to evaluate the models for their performance and better kernel identification. Another set of 16 molecules with ten +1 category and six -1 category molecules are grouped and kept for use as validation data set. Then a set of 53 molecules (from the remaining 45 molecules and 8 molecules taken from validation) is used to constitute the training set data. The details of formation of data sets are given in Table 2.

Table 1. Experimental data set

Drug Category	Molecules	Molecular Count	Total No. of Descriptors
Anticonvulsants	**Camazepam, Primidone , Tofisopam , Beclamide , Fludiazepam , Topiramate, Ketazolam , Nitrazepam, Halazepam , Paraldehyde, Felbamate, Estazolam , Paramethadione , Ethotoin , Cinolazepam , Diazepam, Eslicarbazepine acetate, Etifoxine, Zonisamide, Temazepam, Clomethiazole, Lamotrigine, Fosphenytoin, Carbamazepine, Clotiazepam,**	77	770

	Prazepam, Quazepam, Loprazolam, Meprobamate, Clorazepate, Trimethadione , Phensuximide , Phenacemide, Metocurine, Clonazepam, Aprobarbital, Mephenytoin, Adinazolam, Brotizolam , Metharbital , Methazolamide, Magnesium Sulfate, Stiripentol , Etizolam , Pregabalin, Delorazepam, Levetiracetam, Secobarbital , 2–deoxyglucose, Mexiletine, Alprazolam, Riluzole, Flunitrazepam , Vaproic Acid, Tiagabine, Methsuximide, Lorazepam, Ethosuximide, Vigabatrin, Aminoglutethimide, Acamprosate, Oxcarbazepine, Acetazolamide, Triazolam, Rufinamide, Zolpidem, Chlordiazepoxide, Oxazepam, Clobazepam, Pheobarbital, Phenytoin, Guinidine, Buspirone, Gabapentin , Ergocalciferol , Methylphenobarbital, Dihytrotachysterol		

Table 2. Formation of data sets

Data Sets	Total Molecules	Class-1	Class-II
Training Set*	53	35	18
Validation Set*	16	10	6
Test Set	16	12	4
Entire Set	77	57	28

***** Training set and Validation set data contain descriptor sets of eight molecules in common.

3. Results and Discussion

3.1. Choice of kernels

Radial classification model was tried first. Then polynomial and anova models were tried subsequently. The principle of SVM is supervised learning. Its learning strategy tries to keep the error to the minimum values the method can be reliably used even for scrambled data sets. The computations are performed using, winSVM [7], a Windows implementation of a support vector machine.

During the optimization stage of the training set data, a lower mean square error with 100 percent accurate training phase is searched for the given modelling parameters of C and epsilon. The classification parameter also is looked out carefully to have a value neither too high nor too low in order to avoid over fitting/under fitting of data. The maximum support vectors and bound support vectors available for model training are also watched. Based on these criteria four best performing training models, two for radial and two for polynomial kernels are identified for further validation and testing. The chosen models with the better training parameters identified in the optimization stage are summarized in Table 3. All the six models learnt the training data set completely, evident from the accuracy, precision and recall values being unity. But, in an earlier experiment [8] with radial and polynomial kernels, for antihistamine molecular data set, maximum accuracy was obtained only for polynomial kernel in the validation and independent test phases. But in this experiment, the radial kernel seemed to solve this drug like, non-drug like classification problem better with normalized training set data. To explore more, the efficiency of anova kernel is also studied in this work along with radial and polynomial kernels with direct descriptor data set and normalized data set. To standardize the independent descriptor data set, normalization has been done in the scaling range of 0 to 1. Moreover, such pre-processed data are usually ideal for machine learning processes.

Table 3. SVM training phase parameters

Model Code	Kernel	C	ξ	γ	°	SV	BSV	MSE	Performance Matrix
I	**Radial (with Direct Descriptors)**	**1000**	**0.000001**	**0.2**	**-**	**24**	**0**	**2.47700**	**35 0 0 18**
II	**Radial (with Normalized Descriptors)**	**1000**	**0.000001**	**1**	**-**	**50**	**0**	**0.002048**	**35 0 0 18**
III	**Polynomial (with Direct Descriptors)**	**10000**	**0.01**	**-**	**3**	**18**	**0**	**12.56244**	**35 0 0 18**
IV	**Polynomial (with Normalized Descriptors)**	**0.1**	**0.01**	**-**	**3**	**22**	**1**	**7.25650**	**35 0 0 18**
V	**Anova (with Direct Descriptors)**	**100**	**0.001**	**1**	**3**	**26**	**0**	**0.538176**	**35 0 0 18**
VI	**Anova (with Normalized Descriptors)**	**100**	**0.001**	**1**	**5**	**31**	**0**	**0.13122**	**35 0 0 18**

3.2. Evaluation criteria from classification model

The performance of the SVM models generated for molecular classification has been evaluated by three measures viz. accuracy, sensitivity and specificity. Accuracy is the percentage of correct predictions, for both Class-I and Class-II molecules. Sensitivity is the percentage of Class-I molecules that are correctly predicted as Class-I molecules while specificity is the percentage of Class-II molecules that are correctly predicted as Class-II molecules**.** A statistical parameter called Mathew's Correlation Constant (MCC) is used for the optimization of parameters and to analyse the performance of the model. MCC=1 signifies perfect prediction while MCC= 0 suggests completely random prediction. These evaluation parameters can be represented mathematically as:

Accuracy = {(TP+TN)/(TP + FP+TN+FN)}

Sensitivity or Accuracy on Class-I (Drug-like Molecules) = {TP/ (TP + FN)}

Recall or Precision Class-I (Drug-like Molecules) = {TP/ (TP + FP)}

Specificity or Accuracy on Class-II (non-drugs) = {TN/ (TN + FP)}

Recall or Precision Class-II (non Drug-like Molecules) = {TN/(TN + FN)}

$MCC = [(TP \times TN) - (FN \times FP)] / [\sqrt{[(TP + FN)\ (TN + FP)\ (TP + FP)\ (TN + FN)]}]$

where, True Positive (TP) and True Negative (TN) are correctly predicted Class-I and Class-II molecules, respectively. Similarly, False Positive (FP) and False Negative (FN) are wrongly predicted Class-II and Class-I molecules, respectively. To check whether the present classification scheme is better than a random prediction, a reliability factor (R) can be computed. It is given as

R = [((TP + FN) * (TP + FP)) + (TN + FN) * (TN + FP))] / (TP + TN + FP + FN)

This will give an anticipated number of molecules that could be correctly classified by random prediction. A factor **S** which is independent of the total number of samples in the data set can also be computed as

S = ((TP + TN) - R) / [((TP + TN + FP + FN) - R) x 100]

and it gives the normalized percentage of correctly classified Class-II molecules better than random classification. A value of S = 100% stands for a perfect classification and S = 0% for a poor classification. Apart from these parameters, the overall performance may be revealed out by a static called F1 parameter. It is the harmonic mean of precision and recall (or between sensitivity and positive predictive value). It is given as F1 = (2 x TP) / (2 x TP + FP + FN)

Among these seven parameters, sensitivity, specificity and accuracy can be deemed as basic parameters and the remaining four viz., MCC, R, S and F1 parameters can be regarded as evaluation parameters of best model selection. Based on the better training phase models, the drug like molecular prediction is made with the validation set of data and independent set of data. The computed values of these parameters are listed Tables 4a-4c for the radial, polynomial and anova kernels respectively. A comparative perusal of accuracy, precision and recall parameters reveal the relatively better performance by the radial and anova kernels than the polynomial kernel in both the validation and test phases. The zero MCC value of these to kernels suggest for training with more data. The consistencies in the reliability factor values ensure the fact that these models do not predict the drug likeness of the molecules randomly. However the prediction ability of the models for class-II molecules is appreciable in all the three models. The anova and radial kernels outperformed the rest of the two kernels in the drug likeness prediction, evident from the F1 values. Generally, the best SVM model on the basis of these is to be validated against the test set data not used in the training phase.

In this experiment, a separate external data set has been used for the testing phase and hence the conclusions are agreeable. Also, the important benefit of the support vector machine approach is that the complexity of the resulting classifier is characterised by the number of support vectors rather than the dimensionality of the transformed space. As a consequence, support vector machines tend to be less susceptible to problems of over fitting when compared to other machine learning methods. For bench marking the performance of the radial kernel, this kernel's predictions have been validated with the drug likeness prediction web server tool called Drugmint [9]. The 16 molecules in the external data set have been tested for drug likeness and the results are presented in Table 5.

Table 4a. Evaluation of the performance of the SVM model with Radial Kernel

Kernels/ Parameters	Using Direct Descriptors		Using Normalized Descriptors	
	Validation Phase	Test Phase	Validation Phase	Test Phase
Performance Matrix	8 2 1 5	12 0 2 2	10 0 6 0	12 0 4 0
Accuracy	0.8125	0.875	0.625	0.75
Precision	0.8	0.8571	0.625	0.75
Recall	0.8	1	1	1
MCC	0.70370	1	0	0
R	8.25	11	10	12
S	6.1290	0.006	0	0
F1	0.8421	0.92307	0.769230	0.85714

Table 4b. Evaluation of the performance of the SVM model with Polynomial Kernel

Kernels/ Parameters	Using Direct Descriptors		Using Normalized Descriptors	
	Validation Phase	Test Phase	Validation Phase	Test Phase
Performance Matrix	6 4 3 3	8 4 1 3	5 5 5 1	5 7 1 3
Accuracy	0.5625	0.6875	0.375	0.5
Precision	0.6666	0.8889	0.5	0.8333

Recall	**0.6**	**0.66667**	**0.5**	**0.41666**
MCC	**0.1111**	**0.37037**	**-0.55555**	**0.22222**
R	**8.25**	**8.5**	**8.5**	**7**
S	**9.67741**	**3.3333**	**-3.3333**	**1.111111**
F1	**0.63157**	**0.76190**	**0.5**	**0.55556**

Table 4c. Evaluation of the performance of the SVM model with Anova Kernel

Kernels/ Parameters	**Using Direct Descriptors**		**Using Normalized Descriptors**	
	Validation Phase	**Test Phase**	**Validation Phase**	**Test Phase**
Performance Matrix	**8 2 0 6**	**12 0 0 4**	**10 0 6 0**	**12 0 4 0**
Accuracy	**0.875**	**1**	**0.625**	**0.75**
Precision	**1**	**1**	**0.625**	**0.75**
Recall	**0.8**	**1**	**1**	**1**
MCC	**0.8887**	**1.333**	**0**	**0**
R	**9**	**10**	**10**	**12**
S	**0.0075**	**0.01**	**0**	**0**
F1	**0.8889**	**1**	**0.76923**	**0.85714**

Table 5. A comparison of the predictions of the three different kernels with Drugmint

Sl. No.	**Tested Molecules**	**FT**	**DM**	**RK(D)**	**RK(N)**	**PK(D)**	**PK(N)**	**AK(D)**	**AK(N)**
1	Oxcarbazepine	Y	N	Y	Y	Y	Y	Y	Y
2	Triazolam	Y	Y	Y	Y	Y	N	Y	Y
3	Rufinamide	Y	Y	Y	Y	Y	Y	Y	Y
4	Zolpidem	Y	Y	Y	Y	Y	N	Y	Y
5	Chlordiazepoxide	Y	N	Y	Y	Y	N	Y	Y
6	Oxazepam	Y	N	Y	Y	Y	N	Y	Y
7	Clobazam	Y	Y	Y	Y	Y	Y	Y	Y
8	Phenobarbital	Y	N	Y	Y	N	Y	Y	Y
9	Phenytoin	Y	N	Y	Y	Y	Y	Y	Y
10	Quinidine	Y	N	Y	Y	N	N	Y	Y
11	Buspirone	Y	Y	Y	Y	N	N	Y	Y
12	Gabapentin	N	N	Y	Y	N	N	Y	Y
13	Dihdrotachysterol	N	N	Y	Y	N	N	N	Y
14	Methylphenobarbital	Y	Y	N	Y	Y	Y	N	Y
15	Ergocalciterol	N	N	Y	Y	N	N	N	Y
16	Acetazolamide	N	Y	N	Y	N	N	N	Y

Y – Yes; N –No; FT- Filter tests; DM- Drugmint; D- Direct Descriptors
N- Normalized Descriptors RK- Radial Kernel; PK- Polynomial Kernel; AK- Anova Kernel

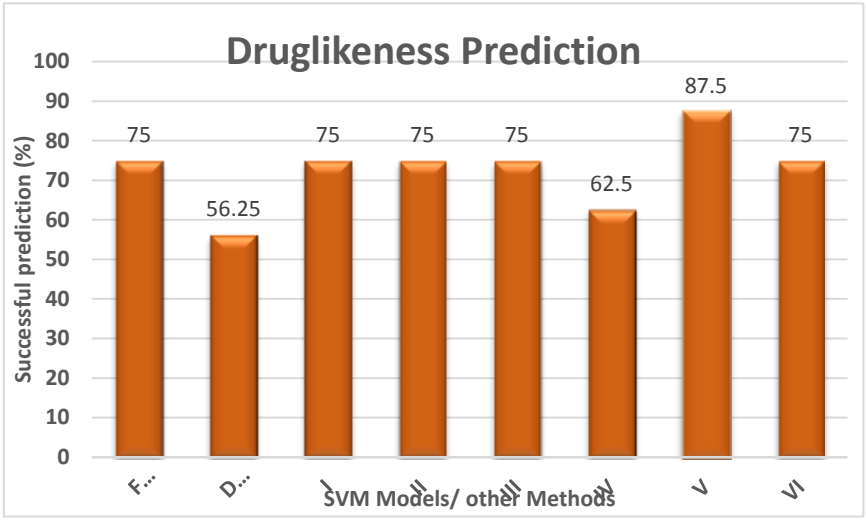

Figure 2. Prediction performance by different SVM models

The comparative performance of the radial, polynomial and anova kernels for each of these 16 molecules is given side by side. The results show a relatively high degree of successful prediction capacity (87.5 %) by anova kernel using direct descriptors (Figure 2). The remaining models show successful prediction rate at 75% or less. The online utility Drugmint's prediction ability of only 56.25 % for this independent data set. It is to be noted that Drugmint also uses SVM score but its learning strategy is based on huge set of molecules derived from different data bases and different types of molecular descriptors but not structure based ones. In the light of the above, the use of structure based descriptors seems to be promising with the help of anova kernel with direct descriptor data in drug likeness prediction of molecules.

Conclusion

Virtual screening of molecules is a regular process in drug development process. In this study, SVM models are generated and evaluated on a series of 77 approved and investigational drug molecules. The best SVM model with total accuracy of 100 % for training set was achieved using a set of 530 structure based descriptors pertaining to 53 molecules. Overall accuracy of up to 87.5 % and 100 % are achieved with the anova kernel (using direct descriptors) in the validation and independent test phases respectively. The findings encourage us to go for more intensive evaluation of the suitability of anova kernel in drug likeness prediction of molecules by experimenting with other categories of molecules. Also it is found that direct descriptors have an edge over normalized ones in training the models and in the successful prediction of drug molecules.

References

[1] R. Todeschini, V. Consonni, Handbook of molecular descriptors. WileyVCH, Weinheim, 2000.

[2] B.E. Boser, I.M. Guyon, V.N. Vapnik. A training algorithm for optimal margin classifiers: D. Haussler (Eds.) In 5th Annual ACM Workshop on COLT, ACM Press, Pittsburgh, PA, 1992, pp. 144-152.

[3] J.S. Taylor, N. Cristianini, Kernel Methods for Pattern Analysis. Cambridge UP, Cambridge, UK, 2004.

[4] B. Scholkopf, A. Smola, A. Learning with Kernels. MIT Press, Cambridge, MA, 2002.

[5] A. Daina, O. Michielin, V. Zoete, SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. Sci. Rep. 7 (2017) 42717.

[6] D.S. Wishart, Y.D. Feunang, A.C. Guo, E.J. Lo, A. Marcu, J.R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, N .Assempour, I. Iynkkaran, Y. Liu, A. Maciejewski, N. Gale, A. Wilson, L. Chin, R. Cummings, D. Le, A. Pon, C. Knox, M. Wilson. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2017 Nov 8.

[7] winSVM. A Windows implementation of a support vector machine based on mySVM software. URL: http://winsvm.martinsewell.com

[8] R.Subbiramanian, T.V.Sundar, G. Vinotha and B.Jeevarathinam (2018). (Communicated)

[9] S.K. Dhanda, D. Singla, A.K. Mondal, G.P. Raghava. DrugMint: a webserver for predicting and designing of drug-like molecules. Biology Direct. 2013;8:28.