Original Article

A prediction model for oral bioavailability of drugs using physicochemical properties by support vector machine

Rajnish Kumar[1,2], **Anju Sharma**[1,2], **Pritish Kumar Varadwaj**[1]

[1]*Department of Bioinformatics, Indian Institute of Information Technology Allahabad, Deoghat, Jhalwa, Allahabad,* [2]*Amity Institute of Biotechnology, Amity University Uttar Pradesh, Lucknow, Uttar Pradesh, India.*

Address for correspondence:
Miss. Anju Sharma, IIITA, Jhalwa, Allahabad, Uttar Pradesh, India. E-mail: anjusharma.online@gmail.com

Abstract

Objective: A computational model for predicting oral bioavailability is very important both in the early stage of drug discovery to select the promising compounds for further optimizations and in later stage to identify candidates for clinical trials. In present study, we propose a support vector machine (SVM)-based kernel learning approach carried out at a set of 511 chemically diverse compounds with known oral bioavailability values. **Material and Methods:** For each drug, 12 descriptors were calculated. The selection of optimal hyper-plane parameters was performed with 384 training set data and the prediction efficiency of proposed classifier was tested on 127 test set data. **Results:** The overall prediction efficiency for the test set came out to be 96.85%. Youden's index and Matthew correlation index were found to be 0.929 and 0.909, respectively. The area under receiver operating curve (ROC) was found to be 0.943 with standard error 0.0253. **Conclusion:** The prediction model suggests that while considering chemoinformatics approaches into account, SVM-based prediction of oral bioavailability can be a significantly important tool for drug development and discovery at a preliminary level.

Key words: Drugs, machine learning, oral bioavailability, prediction, support vector machine

INTRODUCTION

A drug intended for use in humans should have an ideal balance of pharmacokinetics and safety, as well as potency and selectivity. Human oral bioavailability is an important pharmacokinetic property[1] which describes the fraction of an administered drug that reaches the systematic circulation and its site of action, to exert its pharmacological and therapeutic effects. Bioavailability is 100% when a medication is administered parenterally as it goes straight into the bloodstream and is usually completely used by the body. However, when a medication is administered via other routes (such as orally), its bioavailability decreases.

Access this article online	
Quick Response Code:	Website: www.jnsbm.org
	DOI: 10.4103/0976-9668.92325

Prediction of oral bioavailability is not an easy task, as bioavailability depends on superposition of two processes: absorption and liver first-pass metabolism. Absorption in turn depends on solubility and permeability of compounds, as well as interactions with transporters and metabolizing enzymes in gut wall. Permeability further depends on the size of the molecule, as well as its capacity to make hydrogen bonds, its overall lipophilicity and possibly its shape and flexibility. Molecular flexibility, for example, as evaluated by counting the number of rotatable bonds, has been identified as a factor influencing bioavailability.[2-4] The bioavailability of drugs from oral formulations is also influenced by many physiological factors including gastrointestinal fluid composition, pH and dynamics, transit and motility and transport. These factors may vary with age, gender, race, food, and disease.[5] Oral bioavailability is denoted by the letter F.

To lower the attrition rate of drug development there is a need to develop strong and accurate computational methods that can predict and prioritize compounds before they are synthesized or moved towards to preclinical and clinical

development.[6] Various prediction models are reported in the literature on known oral bioavailable drugs such as statistical models,[7-15] mechanistic models,[16-21] QSAR/QSPR models,[22-28] genetic programming,[29-33] artificial neural networks, machine learning classification,[34-36] etc.

MATERIALS AND METHODS

Dataset

We have selected oral bioavailability data from various literature studies.[4,15,37-41] The whole dataset comprises of 1664 drugs. Redundancy was completely removed by manually screening and selected dataset for this study comprises of chemically diverse 511 drugs. Drugs having oral bioavailability less than 30% were regarded as low orally bioavailable drugs and drugs with oral bioavailability 30% or more were regarded as high orally bioavailable.[15] Class labels were defined as "1" for high oral bioavailability and '0' for low oral bioavailability. Further the whole dataset of 511 drug molecules was randomly split into training set of 384 drugs and test set of 127 drugs. Training set was used for training various classifiers, while testing set was not exposed to the system during descriptor selection, learning, kernel selection, and hyper-parameter selection phases.

Descriptor selection

In classification problem usually the data that is to be classified is associated with a large number of features or descriptors. As a result, we get large dimension feature space, making classification a bit difficult task. So first and foremost step is to reduce the dimensions. Feature or descriptor selection is a process of identifying and removing as much of the irrelevant and redundant information as possible. The removal of irrelevant and redundant information often improves the performance of machine learning algorithms. Twelve optimal descriptors were selected using the sequential forward feature selection (SFFS) algorithm.[42]

SFFS algorithm starts with an empty set of features. In first iteration, algorithm considers all feature subsets with only one feature. Feature subset with higher accuracy is used as basis of next iteration. Iteratively algorithm adds to the basis each feature which was not previously selected and retains the feature subset that results in the highest estimated performance. The search terminates after the accuracy of the current subset cannot be improved by adding any other feature. SFFS is stated as: Given a feature set X={xi | i=1…N}, find a subset YM ={xi, …, xM}, with M<N, that optimizes an objective function J(Y).

The set of optimal descriptors include molecular mass (MA), molecular surface area (MSA), molecular volume (MV), molecular refractivity (MR), total hydrogen count (HC), partition coefficient (logP), rotatable bonds (RTB), total polar surface area (TPSA), solubility index (logS), shape flexibility index (SFI) sum of E-states indices (SESI) and count of hydroxyl groups (HYG).

Different feature values for dataset falls in different ranges hence to avoid the discrepancy we have further scaled down these numeric values between –1 to 1. Such scaling facilitates better representation of feature values in kernel function and also avoid numerical difficulties during the calculation.

Support vector machine description

In this process, input vector for training as well as the test set has been quantified as $X^i=(X^i_1, X^i_2,\ldots\ldots, X^i_{13})$, each labeled by corresponding $y^i = 0$ or $y^i = 1$ depending on whether it represents high orally bioavailable drug or low orally bioavailable drug, respectively.

Training set was then subjected to the support vector machine (SVM) classifier, which involved fixing several hyper-parameters which further determines the function optimized by SVM. It is extremely crucial and has a profound impact on the performance of trained classifier. We used several kernels: linear, polynomials, and radial bias function (RBF) initially to determine which of them is applicable to our data and is able to classify it efficiently.[43] We found RBF as the suitable classifier function (as the number of features was not very large in comparison to the dataset) for which training errors on low oral bioavailability data (false negatives) outweigh errors on high oral bioavailability data (false positives).

$$K(X_i, X_j) = \exp(-\gamma \,||\, X_i - X_j \,||^2) \quad \ldots\ldots(1)$$

where $\gamma > 0$.

This kernel (1) is best for the data in which the class-conditional probability distribution function approaches the Gaussian distribution. It maps the non-linear data into a higher dimension space where data is linearly separable. Its exponential nature can be expanded into an infinite series, giving rise to an infinite-dimension polynomial kernel. However, this kernel is bit difficult to design, in the sense that it is difficult to arrive at an optimum "γ" and choose the corresponding C that works best for a given problem. This has been taken care by running grid parameter search exploring all combinations of C and γ with each cross-validation routine, where γ ranged from 2^{-15} to 2^4 and *C* ranged from 2^{-5} to 2^{15}.[44] To identify an optimal hyper-parameter set we have performed a two-step grid-search on C and γ with the use of 10 folds cross-validation, by dividing training set into 10 subsets of equal size (~38

drugs each having 12 descriptors). Iteratively each subset is tested using the classifier trained on the remaining nine subsets. Pairs of (C; γ) have been tried and the one with the best cross-validation accuracy has been picked. Using RBF kernel, the best cross-validation accuracy was obtained at γ = 0.0078125 and C = 512. The result obtained showed a good classification accuracy of 88.54% during the cross-validation. Adopted methodology for model generation is illustrated in Figure 1.

RESULTS

To optimize the SVM parameters γ and *C*, 10-fold cross-validation was applied on each of the training datasets bin, exploring various combinations of *C* (2^{-5} to 2^{15}) and γ (2^{-15} to 2^{4}). In 10-fold cross-validation, the training dataset (384 drugs, each having 12 descriptors) was spilt into 10 subsets, each of equal size, where one of such subsets was used as the test dataset while the other subsets were used for training the classifier. The process is repeated 10 times using a different subset of a corresponding test and training datasets, hence ensuring that all subsets are used for both training and testing. A twofold grid optimization has been considered and the result shown [Figure 2] suggests that the optimized C and γ were found to be 512 and 0.0078125, respectively.

The best combination of γ and C that was obtained from grid based optimization is used for training a RBF-based SVM classifier using entire training data (384 drugs each having 12 descriptors). The result obtained showed a good classification accuracy of 88.54% during the cross-validation. The reported accuracy on the training datasets depicts the effectiveness and reliability of this prediction method; but still it may or may not give the equivalent or better accuracy when applied on the novel drugs, i.e. drugs with an unknown oral bioavailability profile. Therefore, it is extremely important to test the SVM classifier on the non-cross validated test set which is out-of-sample and independent of the training set data. We applied the SVM classifier on the whole test set (127 ligands each having 12 descriptors), the classifier incurred an accuracy of 96.85% by using the RBF kernel with γ = 0.0078125 and C = 512. This prediction accuracy suggests that SVM-based prediction of oral bioavailability can be considered as a helpful tool in drug discovery and development.

The efficiency of a classifier was further evaluated with the help of various quantitative variable: (a) true positive (TP), represents total number of correctly classified high orally bioavailable drugs, (b) true negative or (TN), represents total number of correctly classified low orally bioavailable drugs (c) false positive (FP), represents total number of incorrectly classified low orally bioavailable

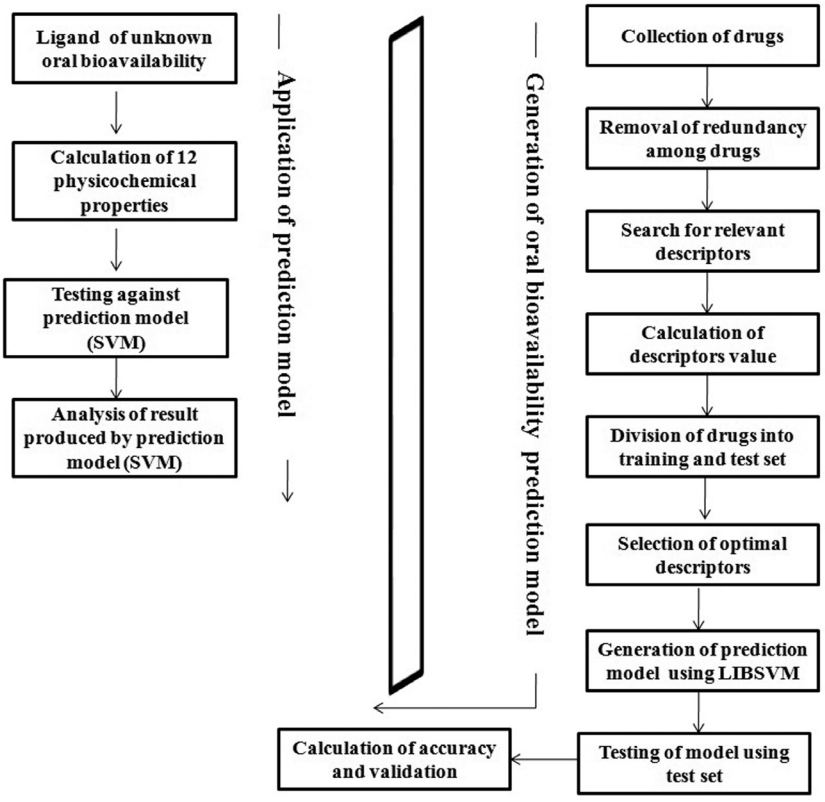

Figure 1: Stepwise illustration of generation of an oral bioavailability prediction model and its application

drugs, (d) false negatives (FN), represents total number of incorrectly classified high orally bioavailable drugs. Using these quantitative variables, several statistical metrics were calculated to measure the effectiveness of the oral bioavailability-SVM classifier.

Sensitivity (Sn) and *specificity (Sp)* metrics, which indicate the ability of a prediction system to classify the high and low orally bioavailable drugs, were calculated by equations (2) and (3) and receiver operating characteristic curve (ROC) for the same was plotted [Figure 3].

$$Sn\ (\%) = [TP/\ (TP+FN)]*100 \quad .....(2)$$

$$Sp\ (\%) = [TN/(TN+FP)*100] \quad .....(3)$$

To indicate an overall performance of the classifier system; accuracy (Ac), for the percentage of correctly classified drugs and the Matthews correlation coefficient (MCC) were computed as follows:

$$Ac = [(TP + TN)/\ (TP+FP+TN+FN)]*100 \quad .....(4)$$

$$MCC= [(TP*TN)-(FP*FN)]/\sqrt{\ (TN+FP)\ (TN+FN)\ (TP+FP)\ (TP+FN)} \quad .....(5)$$

Sensitivity (Sn) came out to be 95.60% with a false positive proportion (FP) of 0.79% whereas specificity (Sp) came out to be 97.30% with a false negative (FN) proportion of 3.15%. Similarly Youden's Index (Youden's Index = sensitivity + specificity – 1) was 0.929 and Matthews correlation coefficient (MCC) was found to be 0.909. The overall accuracy (Ac) calculated using equation (4) was 96.1% which is significantly higher than existing methods. The area under ROC curve was found to be 0.943 with a standard error of 0.0253.

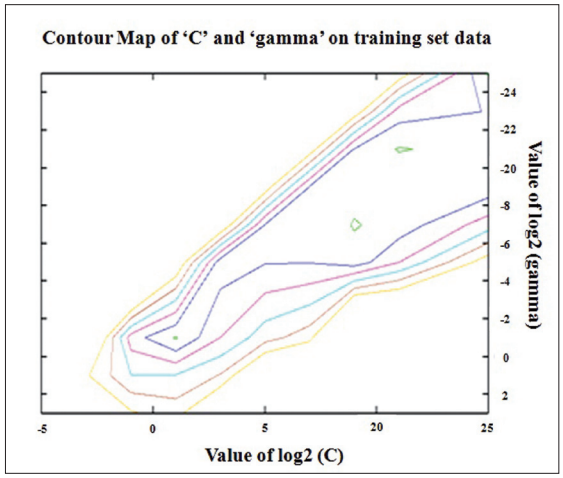

Figure 2: Contour plot of grid search result showing optimum values of hyper-parameter

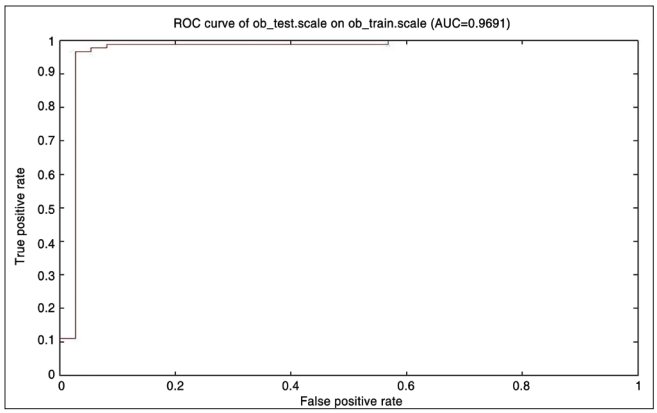

Figure 3: Receiver operating characteristic (ROC) plot for a classifier with optimized values of C and γ

DISCUSSION

The prediction model derived from SVM can serve as primary tool for generating some idea about oral bioavailability of ligands. User just needs to calculate the 12 physicochemical descriptors, as these values are prerequisite for prediction of oral bioavailability through the generated SVM model [Figure 1]. The ligand with unknown oral bioavailability can be tested against the prediction model. For given 12 physicochemical properties this SVM model can predict the oral bioavailability of the ligand under consideration. At preliminary level, this model can predict that whether the oral bioavailability of the ligand under study is low or high.

Numerous attempt have been made to predict oral bioavailability of drugs and ligands by computational and experimental method in past. Some of those prediction models are listed in table 1 along with the current study and the model generated by SVM seems to be more satisfactory in terms of prediction accuracy.

Absorption of drug taken orally is a complex process and, although related to drug physicochemical properties, it is related in fairly complex ways. Physiological and environmental conditions influence the bioavailability of drugs such as the presence or the absence of food, residence time of the drug in contact with the small intestinal epithelium, etc and make the absorption prediction further complex.[45] Failure to appreciate this complexity in attempting to build models may lead to the generation of model with low confidence. An alternative approach to modeling oral bioavailability is to develop structure-based models for the properties contributing to the absorption process, such as solubility and permeability (included in presented model as logS and logP). These can then be used to identify opportunities for optimization. For example, if a potential drug is expected to have poor

Table 1: Comparative study of some of the oral bioavailability prediction models with current study

Oral bioavailability prediction methods	Size of dataset	Accuracy (%)	Citation
Statistical approach	1000	85	26
Mechanism-based approach (Gastroplus software)	50	80	46
QSAR/QSPR approach (fuzzy adaptive least squares)	232	67	22
QSBR model	1261	79-86	25
Genetic programming-based approach (GA-CG-SVM)	690	86	47
SVM approach	511	96.85	Present study

oral bioavailability due to low-intrinsic aqueous solubility, then this is a property amenable to manipulation by the formulation scientist. On the other hand, if the compound is both poorly soluble and permeable, along with a significant metabolic liability, optimization may be very difficult if not impossible. Such candidates present high risks to successful development and should be identified as such early in the drug identification and development process. Judicious development and use of computational models will clearly aid in these processes.[46,47]

CONCLUSION

The SVM classifier with radial basis function kernel with $\gamma = 0.0078125$ and $C = 512$ applied on the test datasets. The overall accuracy of the model obtained is 96.85%. It suggests that while considering chemoinformatics approaches into account, SVM-based prediction of oral bioavailability can be a significantly important tool for drug development and discovery at a preliminary level.

REFERENCES

1. Moda TL, Montanari CA, Andricopulo AD. Hologram QSAR model for the prediction of human oral bioavailability. Bioorg Med Chem 2007;15:7738-45.
2. Waterbeemd H, Gifford E. ADMET in silico modeling: Towards prediction paradise? Nat Rev Drug Disc 2003;2:192-204.
3. Van de Waterbeemd H, Rose S. In The Practice of Medicinal Chemistry. 2nd ed. In: Wermuth LG, editor. Academic Press; 2003. p. 1367-85.
4. Veber DF, Johnson SR, Cheng HY, Smith BR, Ward KW, Kopple KD. Molecular properties that influence the oral bioavailability of drug candidates. J Med Chem 2002;45:2615-23.
5. Masoud J, Dabid T, Jiansong Y, Sibylle N, Sebastian P, Amin RH, *et al*. Population based mechanistic prediction of oral drug absorption. AAPS J 2009;11:225-37.
6. Stephen RJ, Weifan Z. Recent progress in the computational prediction of aqueous solubility and absorption. AAPS J 2006;8: E27-40.
7. Lipsinki CA, Lomabardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Delivery Rev 2001;46:3-26.
8. Manuel AN, Pravin RC. Design principles for orally bioavailable drugs. Drug Discov Today 1999;1:179-89.
9. Hirschmann R. Peptide research a means to further biological and chemical understanding. In Peptides 1996. Proceedings of the 24th European peptide Symposium.
10. Martin TK, Yvonne A. Influence of molecular properties on oral bioavailability of lipophilic drugs-Mapping of bulkiness and different measures of polarity. Pharm Dev Technol 2009;14:312-20.
11. Lu JJ, Crimin K, Goodwin JT, Crivori P, Orrenius C, Xing L, *et al*. Influence of molecular flexibility and polar surface area metrics on oral bioavailability in the rat. J Med Chem 2004;47:6104-7.
12. Martin YC. A bioavailability score. J Med Chem 2005;48:3164-70.
13. Vieth M, Siegel MG, Higgs RE, Watson IA, Robertson DH, Savin KA, *et al*. Characteristic physical properties and structural fragments of marketed oral drugs. J Med Chem 2004;47:224-32.
14. Sietsema WK. The absolute oral bioavailability of selected drugs. Int J Clin Pharmacol Ther Toxicol 2000;44:235-49.
15. Hou T, Wang J, Zhang W, Xu X. ADME evaluation in Drug Discovery.6. Can Oral bioavailability in humans be predicted by simple molecular property-based rules. J Chem Inf Model 2007;47:460-3.
16. Somogyi A, Eichelbaum M, Gugler R. Prediction of bioavailability for drugs with a high first-pass effect using oral clearance data. Eur J Clin Pharm 2006;22:85-90.
17. Usansky HH, Sinko PJ. Estimating human drug oral absorption kinetics from Caco-2 permeability using an absorption-disposition model: Model development and evaluation and derivation of analytical solutions for k(a) and F(a). J Pharmacol Exp Ther 2005;314:391-9.
18. Mahmood I. Prediction of absolute bioavailability for drugs using oral and renal clearance following a single dose: A critical view. Biopharm and Drug Dispos 1997;18:465-73.
19. Obata K, Sugano K, Saitoh R, Hiqashida A, Nabuchi Y, Machida M, *et al*. Prediction of oral drug absorption in humans by theoretical passive absorption model. Int J Pharm 2005;293:183-92.
20. Willmann S, Schmitt W, Keldenich J, Lippert J, Dressman JB. A physiological model for the estimation of the fraction dose absorbed in humans. J Med Chem 2004;47:4022-31.
21. Arun KM, Thomas NT, Hwang KK. Graphical model for estimating oral bioavailability of drugs in humans and other species from there Caco-2 permeability and in vitro liver enzyme metabolic stability rates. J Med Chem 2002;45:304-11.
22. Yoshida F, Topliss JG. QSAR model for drug human oral bioavailability. J Med Chem 2000;43:2575-85.
23. Podlogar BL, Muegge I, Brice LJ. Computational methods to estimate drug development parameters. Curr Opin Drug Disc Dev 2001;4:102-9.
24. Andrews CW, Bennett L, Yu LX. Predicting human oral bioavailability of a compound: Development of a novel quantitative structure-bioavailability relationship. Pharm Res 2000;17:639-44.
25. Bai JP, Utis A, Crippen G, He HD, Fischer V, Tullman R, *et al*. Use of classification regression tree in predicting oral absorption in humans. J Chem Inf Comput Sci 2004;44:2061-9.
26. Zmuidinavicius D, Didziapetris R, Japertas P, Avdeef A, Petrauskas A. Classification structure-activity relations (C-SAR) in prediction of human intestinal absorption. J Pharm Sci 2003;92:621-33.
27. Joseph VT, Beverly DG, Snezana AK. Prediction of drug bioavailability based on molecular structure. Anal Chimica Acta 2003;485:89-102.
28. Klopman G, Stefan LR, Saiakhov RD. ADME evaluation, II: A computer model for the prediction of intestinal absorption in humans. Eur J Pharm Sci 2002;17:253-63.
29. Francesco A, Stefano L, Enza M, Leonardo V. Genetic programming for human bioavailability of drugs. GECCO'06 2006. p. 255-62.

30. Sara S, Leonardo V. Operator equalization, bloat and overfitting: A study on humab bioavailability prediction. ACM 2009. p. 1115-22.
31. Bains W, Gilbert R, Sviridenko L, Gascon JM, Scoffin R, Birchall K, *et al*. Evolutionary computational methods to predict oral bioavailability QSPRs. Curr Opin Drug Disc Dev 2002;5:44-51.
32. Wang J, Krudy G, Xie XQ, Wu C, Holland G. Genetic algorithm-optimized QSPR models for bioavailability, protein binding, and urinary excretion. J Chem Inf Model 2006;46:2674-83.
33. Pintore M, Van de Waterbeemd H, Piclin N, Chrétien JR. Prediction of oral bioavailability by adaptive fuzzy partitioning. Eur J Med Chem 2003;38:427-31.
34. Turner JV, Maddalena DJ, Agatonovic-Kustrin S. Bioavailability prediction based on molecular structure for a diverse series of drugs. Pharm Res 2004;21:68-82.
35. Fröhlich H, Sieker F, Wegner K, Zell A. Kernel functions for attributed molecular graphs - a new similarity based approach to ADME prediction in classification and regression. QSAR Comb Sci 2005;25:317-26.
36. Liu HX, Hu RJ, Zhang RS, Yao XJ, Liu MC, Hu ZD, *et al*. The prediction of human oral absorption for diffusion rate-limited drugs based on heuristic method and support vector machine. J Comput Aided Mol Des 2005;19:33-46.
37. Zhao YH, Abraham MH, Le J, Hersey A, Luscombe CN, Beck G, *et al*. Rate limited steps of human oral absorption and QSAR studies. Pharmaceut Res 2002;19:1446-57.
38. Oprea I, Gottfries J. Toward minimalistic modeling of oral drug absorption. J Mol Graphics Modell 1999;17:261-74.
39. Kenneth ET, Danny DS. Percentage oral availability. Book: Goodman and Gilman's The Pharmacological Basis of Therapeutics". 10/e, ed. In: Hardman JG, Limbird LE, Gilman AG, editors. McGraw-Hill, p. 1917-2023.
40. Dorronsoro I, Chana A, Abasolo MI, Castro A, Gil C, Stud M, *et al*. CODES/neural network model: A useful tool for in silico prediction of oral absorption and blood-brain barrier permeability of structurally diverse drugs. QSAR Comb Sci 2004;23:89-98.
41. Linnankoski J, Makela JM, Ranta VP, Urtti A, Yliperttula M. Computational prediction of oral drug absorption based on absorption rate constants in humans. J Med Chem 2006;49:3674-81.
42. Sharma A, Kumar R, Varadwaj P. Prediction of mutagenicity of compounds by support vector machine. Online J Bioinfo 2011; 12:9-17.
43. Bennett KP, Campbell C. Support vector machine: Hype or hallelujah? SIGKDD Explor 2000;2:1-13.
44. Chang CC, Lin CJ. LIBSVM: A library for support vector machines. 2001. Software Available from: http://www.csie.ntu.edu.tw/~cjlin/libsvm. [cited in 2001].
45. Burton PS, Goodwin JT, Vidmar TJ, Amore BM. Predicting drug absorption: How nature made it a difficult problem. J Pharmacol Exp Ther 2002;303:889-95.
46. De Buck SS, Sinha VK, Fenu LA, Gilissen RA, Mackie CE, Nijsen M. The prediction of drug metabolism, tissue distribution and bioavailability of 50 structurally diverse compounds in rat using mechanism-based ADME prediction tools. Drug Metab Dispos 2007;35:649-59.
47. Ma CY, Yang SY, Zhang H, Xiang ML, Huang Q, Wei YQ, Prediction models of human plasma protein binding rate and oral bioavailability derived by using GA-CG-SVM method. J Pharmaceut Biomed Anal 2008;47:677-82.

How to cite this article: Kumar R, Sharma A, Varadwaj PK. A prediction model for oral bioavailability of drugs using physicochemical properties by support vector machine. J Nat Sc Biol Med 2011;2:168-73.

Source of Support: Ministry of Human Resource and Development (MHRD), Govt. of India for their financial support at IIIT- Allahabad. **Conflict of Interest:** None declared.