Research Article

Regression Analysis: Identifying Molecular Descriptors for HIA, MDCK and Caco-2

Mukunthan KS[a,c*], Amritendu Bhattacharya[b], Trupti Navin Chandra Patel[c]
[a]Department of Biotechnology, Manipal Institute of Technology, Manipal, Karnataka, India.
[b] Department of Mechanical Engineering and Mathematical Sciences, Oxford Brookes University, United Kingdom.
[c]Division of Medical Biotechnology, School of Bioscience and Technology, VIT University, Vellore, Tamil Nadu, India.
***Corresponding author's E-mail:** mukunthanselvam@gmail.com

Accepted on: 10-02-2016; **Finalized on:** 29-02-2016.

ABSTRACT

Oral bioavailability depends on many physiological, physiochemical and formulation factors. Poor oral bioavailability is an important parameter accounting for the failure of the drug candidates. 50% of drug failure is because of unfavorable bioavailability. Two important properties that govern oral absorption are *in vitro* permeability and solubility, which are commonly used as indicators of Human Intestinal Absorption (HIA) and Colon epithelial cancer cell line (Caco-2) and Madin-Darby Canine Kidney cells (MDCK) for permeability. *In silico* prediction of oral bioavailability based on physiochemical properties are highly needed. Although many computational models have been developed to predict absorption and permeability, their accuracy remains low with a significant number of false positives. In this study, we present model based on systems biological approach, using regression analysis of predictions coupled with physiochemical descriptors. A large dataset of HIA, Caco2, MDCK predictions was collated along with physiochemical descriptors for the chosen chemical structures. The descriptors found common in three regression analysis showed good relation with rule of five descriptors. Nevertheless, the study captures the fundamental molecular descriptors, which can be used as an entity to facilitate increase in oral bioavailability.

Keywords: Bioavailability, HIA, MDCK, Caco-2, Regression analysis, Chemical descriptors, Dragon

INTRODUCTION

The ability to deliver drugs orally is strongly preferred over alternative routes for systemic administration. The administered complex drug molecule is simplified and circulated inside human body by blood. Excess drug molecules after absorption & distribution will be excreted from the body. Any promising chemical molecule during early stages of drug discovery is named as lead. Such lead molecules when they passes through absorption, distribution, metabolism, excretion, and toxicity (ADMET) standards in animal models they are elevated as drug candidates. Almost 40% of drug candidate attrition are caused by adverse pharmacokinetics (PK) and bioavailability[1]. Poor absorption is a major factor that leads to drug attrition. The absorption process involves the entry of drug into systemic circulation from the site of administration. Moreover, the drug's pharmacokinetic profile can be easily and significantly changed by adjusting factors that affect absorption[2]. The oral dosing of drug are calculated using absorption models like: HIA for understanding adsorption of molecules in intestinal wall, Caco-2 for assessing paracellular movement of molecules across the monolayer cells and MDCK for assessing membrane permeability properties of molecules.

"Fail early and fail fast" is the current paradigm that the pharmaceutical industry has adopted widely. Removing non-drug-like compounds from the drug discovery lifecycle in the early stages can lead to tremendous savings of resources[3]. Today, early characterization of drug properties by the computational methods has attracted significant attention in pharmaceutical discovery and development. Scientists use computational methods to screen molecules with reasonable ADMET properties for drug testing during initial phases[4]. Such lead molecules from virtual libraries can then be synthesized and subjected to high-throughput biological activity screening. As the predictive ability of software improves, the drug discovery process will move from a screening-based to a knowledge-based paradigm economically.

In vivo and *Ex vivo* model studies of absorption are labor intensive and variable results. *In silico* models available for the prediction of oral absorption are having high degree of accuracy[5]. Multivariate approaches like multiple linear regression, partial least squares and artificial neural networks, have been used to develop chemical structure and absorption relationships[6].

Among the available computational drug discovery technology, quantitative structure-activity relationship approaches that rely on chemistry descriptors are the most appropriate to design drugs[7]. The models are based on a single descriptor, such as log P or log D, or polar surface area, which is a descriptor of hydrogen-bonding potential to a variety category are employed in drug designing[8]. A combined analysis like clustering and multiple linear regression analysis were now employed to determine the distinct group of molecular descriptors that largely account for the biological activity[9].

In computational chemical modeling, feature selection is used, to reduce the number of descriptors used per chemical molecule. If molecules are represented by improper descriptors, they will not lead to good predictions. Successful drug molecule depends on identification of good descriptor while data mining. The aim of this paper is to identify the chemical descriptors that contribute to better absorption properties.

MATERIALS AND METHODS

The 103 chemical structures of '*Curcuma caesia* Roxb' were collected from literature and experiment was used in this study[10-14]. The structure of all compounds were drawn by using ACD/Chemsketch and converted into molfile (*.mol). The PreADMET program was accessed at http://preadmet.bmdrc.org/.

The program automatically calculated the predictive absorption and permeation for HIA, Caco-2 and MDCK[15].

All structures were used to calculate 1354 molecular descriptors such as Constitutional descriptors, 2 Ring descriptors, Topological indices, Walk and path counts, Connectivity indices, Information indices, 2D matrix-based descriptors, 2D autocorrelations, Burden eigenvalues, PVSA-like descriptors, ETA indices, Edge adjacency indices, Geometrical descriptors, 3D matrix-based descriptors, 3D autocorrelations, RDF descriptors, 3D-MoRSE descriptors, WHIM descriptors, GETAWAY descriptors, Randic molecular profiles, Functional group counts, Atom-centered fragments, Atom-type E-state indices, CATS 2D, 2D Atom Pairs, 3D Atom Pairs, Charge descriptors, Molecular properties, Drug-like indices descriptors using Dragon[16].

Statistical calculations were carried out using the SAS for simple multiple linear regression benefits from a well-developed mathematical framework that yields unique solutions and exact confidence intervals for regression coefficients[17]. Regression analysis was used to identify the significant descriptors using the same molecular descriptors with HIA, Caco2 and MDCK absorption values of the chemical structures.

RESULTS AND DISCUSSION

It was feasible to achieve the required robustness for an *In silico* study based on a relatively large samples under a careful statistical method. Prediction of chemical descriptors aiding to absorption parameter by multiple regression methods using a combination of structure-based molecular descriptors and some absorption models as predicting variables. The descriptors identified for HIA, MDCK and Caco-2 are listed in the **Table 1.** The models R-Square values of HIA, MDCK and Caco-2 was observed as following 0.91, 0.75 and 0.92.

Similar good correlation with R-sq value 0.78 was obtained from the physiochemical descriptors in comparison to structural descriptor calculation in curcumin analogues[18].

The descriptors which are common among the three absorption models are Mwt, HBDH, MNO, TPSA, SlogP, SlogD and MlogP1. The identified descriptors are identical to Lipinski's established useful guidelines for achieving acceptable oral exposure as part of the 'rule of 5'[19-20]. A low p-value which is less than 0.05 indicates the rejection of the null hypothesis. The descriptors chosen with p value are shown in **Table 2.**

The regression equation obtained after HIA analysis

Y = 99.93 + BEHe3 (-8.7) + BEHm2 (5.91) + H-047 (0.27) + MATS4m (-3.83) + Neoplastic-80 (-2.12) + R4u (-59.15) + piPC06 (0.737)

The regression equation obtained for MDCK analysis

Y = 232.15 + CIC4 (54.96) + C_016 (13.85) + EEig04x (-106.22) + G1m (-360.89) + G2s (278.50) + G2v (-462.31) + GATS3e (-106.33) + GNar (164.35) + GVWAI-80 (44.78) + Infective-50 (51.64) + MATS5p (-97.79) + MATS6p (-43.12) + Mor30u (-62.60) + Mor32m (-139.56)

The regression equations obtained for Caco-2 analysis

Y=112.33 + T-PSA (-0.76) + GATS2e (7.73) + GVWAI_50 (-13.15) + Hypnotic_80 (-5.88) + Inflammat_80 (21.23) + MATS6p (-8.04) + Mor13e (9.42) + Mor13m (-12.01) + Mor21v (16.44) + Mor22u (10.33) + P2u (-75.43) + PJI2 (28.34) + R1p (-336.52) + RDF040m (1.20) + RDF090v (-3.47) + RDF095u (-1.19) + X2Av (-202.61) + nCrt (-2.36) + nR06 (-6.90)

Table 1: Selected molecular descriptors for HIA, MDCK and Caco-2

Absorption parameter	Molecular descriptor identified
HIA	Mwt, HBDH, MNO, TPSA, SlogP, SlogD, MlogP1, BEHe3, BEHm2, BELv4, Du, Dv, GVWAI50, GNO, H047, MATS4m, Mor09m, Neoplastic80, R4u, Tl1, nROH and piPC06
MDCK	Mwt, HBDH, MNO, TPSA, SlogP, SlogD, MlogP1, CIC4, C016, EEig04x, G1m, G2s, G2v, GATS3e, GNar, GVWAI80, IC0, Infective0, MATS5p, MATS6p, Mor30u and Mor32m
Caco-2	Mwt, HBDH, MNO, TPSA, SlogP, SlogD, MlogP1, BELp2, GATS2e, GATS8e, GVWAI50, Hypnotic80, inflammat80, JGI5, MATS6p, Mor13e, Mor13m, Mor21v, Mor22u, O058, P2u, PJI2, R1p, RDF040m, RDF090v, RDF095u, X2Av, nCrt and nR06

Table 2: Chemical descriptors with p value cutoff of < 0.05

Absorption parameter	Molecular descriptors select with p value	Descriptors classification
HIA	BEHe3, BEHm2, H047, MATS4m, Neoplastic80, R4u, piPC06	electronegativity-Burden eigen values, atomic mass-Burden eigen values, atomic mass-2D autocorrelations, antineoplastic-molecular properties, Gateway descriptor, molecular multiple path count-walk and path descriptor
MDCK	CIC4, C016, EEig04x, G1m, G2s, G2v, GATS3e, GNar, GVWAI80, Infective50, MATS5p, MATS6p, Mor30u, Mor32m	neighborhood symmetry- information indices, =CHR-atom centered fragments, Matrix weight by edge degree- edge adjacency index, weight by atomic mass - WHIM descriptor, weight by electro topological state - WHIM descriptor, weight by Vander walls value- WHIM descriptor, electronegativity-2D autocorrelation, geometric topology index-topological descriptor, drug like index-molecular properties, anti-infective index- molecular properties, weighted by atomic polarization-2D autocorrelation, Atom mass- 3D MoRSE descriptor.
Caco-2	TPSA, GATS2e, GVWAI50, Hypnotic80, Inflammat80, MATS6p, Mor13e, Mor13m, Mor21v, Mor22u, P2u, PJI2, R1p, RDF040m, RDF090v, RDF095u, X2Av, nCrt, nR06	sanderson electronegativity - 2D autocorrelations, drug like index - Molecular properties, Ghose - Viswanathan - Wendoloski hypnotic - like index- molecular properties, Inflammation- molecular properties, atomic polarizabilities - 2D autocorrelations, electronegativity's - 3D - MoRSE descriptor, atomic masses - 3D - MoRSE descriptor, atom Vander walls volume - 3D - MoRSE descriptor, shape directional - WHIM descriptor, shape index - topological descriptor, atomic polarizabilities - gateway descriptor, atom mass - RDF descriptor, vanderwalls - RDF descriptor, radial distribution - RDF descriptor, average valence connectivity-connectivity index, Number of ring tertiary-functional group count, no of 6 membered rings-constitutional descriptor

The four assumptions of regression analysis Linearity, Independence, Normal distribution and Equal variance were met during the analysis. **Figure 1, 2, 3** shows various plots of HIA, MDCK and Caco-2 parameters.

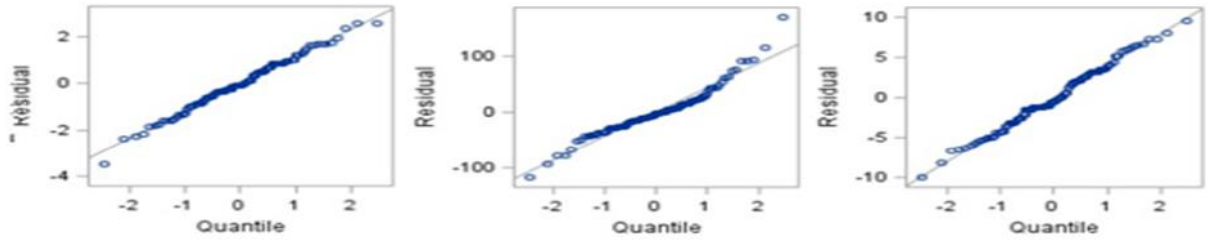

Figure 1 Normality plot

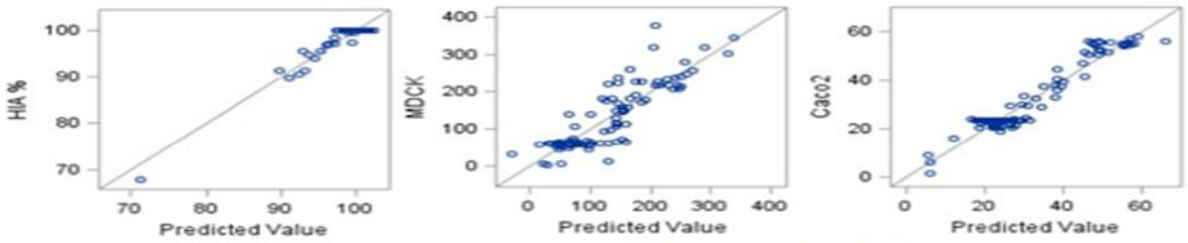

Figure 2 Linearity plot

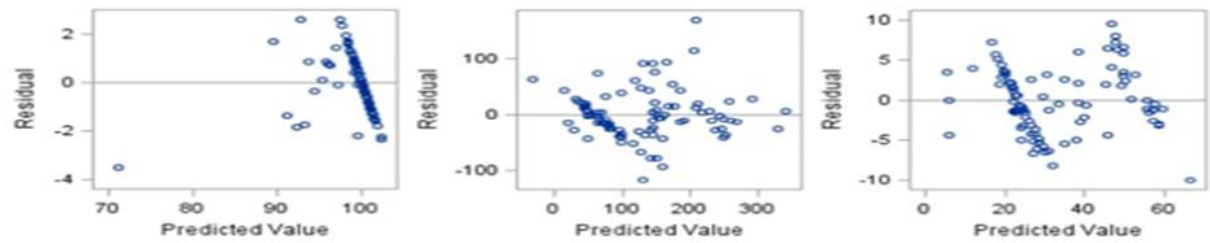

Figure 3 Equality of variance

Oral bioavailability of drugs is strongly influenced by solubility and permeability parameters for their absorption via passive diffusion[21]. The size and shape of the drug molecule will affect absorption. Lesser molecular weight drugs are absorbed better compared to larger ones. As molecular size increases, a larger cavity must be formed in water to soluble. Increasing size also impedes passive diffusion through the tightly packed aliphatic side chains of the lipid bilayer membrane. The smaller drug molecules are better, because diffusion is directly affected. Much of the drugs have molecular weights under 450 daltons and are grouped as small molecules.

Hydrogen bonding is an important parameter for describing drug permeability. Solubility of drug in water can be estimated from the number of hydrogen bond donors against the alkyl side chains in the drug. Low water solubility will lead to slow absorption in intestines. H-bonding capacity of a drug has been found to correlate well with intestinal absorption[22].

Studies proved that intestinal permeability using the hydrogen bonding donors are found to be important than the hydrogen bond acceptors[23]. The presence of many hydrogen bond donors, on the other hand, leads to less penetration of the cell membrane.

The formation of intermolecular hydrogen bonds in drug molecules is also identified as to improved membrane permeability and intestinal absorption in rat and human[24].

Partition coefficient is useful in estimating the difference in solubility of the compound in octanaol and water bi phases. Hydrophobic drugs with high octanol/water partition coefficients are preferentially distributed to hydrophobic compartments such as the lipid bilayers of cells while hydrophilic drugs (low octanol/water partition coefficients) preferentially are found in aqueous compartments such as blood serum[25].

Octanol-water partition coefficient is used as significant tool to measure of molecular hydrophobicity. The hydrophobicity nature of drug will affect drug absorption, bioavailability, hydrophobic drug-receptor interactions and metabolism of molecules, as well as its toxicity[26].

Molecular Topological Polar Surface Area (TPSA) is the sum of surface contributions of polar atoms in a molecule, which shows good correlation with drug transport properties, such as intestinal absorption, or blood-brain barrier penetration[27].

Our results are in good agreement with prediction studies on Caco-2 cellular permeation factors depending on partition coefficient, polar surface area, and radius of gyration. Such identified descriptors included with optimum solubility, H-bonding, and bulk properties in the prediction models permit the interpretation in structural terms of the permeability process[28].

Similar approach was employed to predict MDCK cell permeation coefficients of organic molecules using membrane-interaction QSAR analysis. The most important descriptor identified in the models is ClogP which was in good agreement with our results[29].

The p value cut of found descriptor like electronegativity and bond polarity go hand in hand to explain many of the properties of drug that are crucial for their absorption in human body. Oral absorption efficiency in humans was predicted based on their ionization energy and electronegativity[30-31].

The vander wall interactions between drug molecules and surfaces are critical for the study of adsorption. Vander wall interactions are not limited to protein targets but aids in easy passage of drug during absorption. benzene, hydrocarbon demonstrate high degree of absorption. Lipid soluble structures like steroid nucleus and halogen groups are structures if incorporated in drug will better the absorption[32]. From the discussion we concluded that the molecular weight, hydrogen bond donor, Topological polar surface area, solubility, vander waal, ionization, ring numbers and electronegativity descriptors of the molecules play major role in the absorption and permeation of drug molecules.

CONCLUSION

In conclusion, the systems biology approach was used on HIA, MDCK and Caco-2 data set to determine the major contributing descriptors for oral bioavailability prediction. Overall, 7 descriptors were identified as common between HIA, MDCK and Caco-2 permeability, and it is validated to be crucial in predicting oral absorption. The predictions on data sets demonstrate that this model has good in estimating the oral absorption. The selected descriptors capture Lipinski's features of oral bioavailability and can predict oral bioavailability with accuracy. Overall, this study shows that the choices of both machine learning and optimal descriptor sets are critical for the prediction tasks. Conceivably, a similar approach can be used for the prediction of most contributing descriptors involved in drug distribution and toxicity.

REFERENCES

1. Kola I, Landis J, Can the pharmaceutical industry reduce attrition rates? Nature Reviews Drug Discovery, 3, 2004, 711-716.
2. Hou T, Wang J, Structure-ADME relationship: still a long way to go? Expert Opinion on Drug Metabolism & Toxicology, 4, 2008, 759-0770.
3. Cheng A, Merz K.M Jr, Prediction of aqueous solubility of a diverse set of compounds using quantitative structure-property relationships, Journal of Medicinal Chemistry, 46, 2003, 3572-3580.
4. Xu J, Hagler A, Chemoinformatics and Drug Discovery, Molecules, 7(8), 2002, 566-600.
5. Van de Waterbeemd H, Gifford E, ADMET *in silico* modelling: towards prediction paradise? Nature Reviews Drug Discovery, 2(3), 2003, 192-204.

6. Agatonovic-Kustrin S, Beresford R, Yusof APM, Theoretically-derived molecular descriptors important in human intestinal absorption, Journal of Pharmaceutical and Biomedical Analysis, 25, 2001, 227-237.
7. De Benedetti PG, Fanelli F, Computational quantum chemistry and adaptive ligand modeling in mechanistic QSAR, Drug Discov Today, 15, 2010, 859-866.
8. Tetko IV, GasteigerJ, Todeschini R, Mauri A, Livingstone D, Ertl P, Palyulin VA, Radchenko EV, Zefirov NS, Makarenko AS, Tanchuk VY, Prokopenko VV, Virtual computational chemistry laboratory - design and description, Journal of Computer-Aided Molecular Design, 19, 2005, 453-463.
9. Vios VLS, Billones JB, Cluster and multi-linear regression analyses guided identification of molecular descriptors that account for cyclooxygenase activities, Journal of Chemical and Pharmaceutical Research, 7(8), 2015, 735-742.
10. Banerjee AK, Kaul VK, Nigam SS, Chemical examination of the essential oil of Curcuma caesia Roxb. Essenze e Derivati Agrumari. 54, 1984, 117-121.
11. Buddhasukh D, Smith J, Ternai B, Essential oil of Curcuma caesia Roxb. Journal of the Science Faculty of Chiang Mai University, 21-22(1-2), 1995, 14-16.
12. Pandey AK, Chowdhury AR, Volatile constituents of the rhizome oil of Curcuma caesia Roxb. from central India, 18, 5, 2003, 463-465.
13. Liu Y, Roy SS, NebiÃ© RC, Zhang Y, Nair MG, Functional Food Quality of Curcuma caesia, Curcuma zedoaria and Curcuma aeruginosa Endemic to Northeastern India Plant Foods for Human Nutrition, Plant Foods for Human Nutrition, 68(1), 2013, 72-77.
14. Mukunthan KS, Kumar NVA, Balaji S, Trupti NP, Analysis of Essential Oil Constituents in Rhizome of Curcuma caesia Roxb. from South India Journal of Essential Oil Bearing Plants, 17, 2014, 647-651.
15. Lee SK, Lee IH, Kim HJ, Chang GS, Chung JE, The PreADME Approach: Web-based Program for Rapid Prediction of Physico-Chemical, Drug Absorption and Drug-Like Properties. Blackwell Publishing Massachusetts, 2003, 418-420.
16. Tetko IV, Gasteiger J, Todeschini R, Mauri A, Livingstone D, Ertl P, Palyulin VA, Radchenko EV, Zefirov NS, Makarenko AS, Tanchuk VY, Prokopenko VV, Virtual computational chemistry laboratory - design and description, Journal of Computer-Aided Molecular Design, 19, 2005, 453-463.
17. Satpathy R, Guru RK, Behera R, Computational QSAR analysis of some physiochemical and topological descriptors of Curcumin derivatives by using different statistical methods, Journal of Chemical and Pharmaceutical Research, 2(6), 2010, 344-350.
18. Lipinski CA, Lead- and drug-like compounds: the rule-of-five revolution, Drug Discovery Today: Technologies, 1, 2004, 337-341.
19. Lipinski CA, Franco L, Dominy BW, Feeney PJ, Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings, Advance drug discovery review, 23, 1997, 3-25.
20. Shugarts S, Benet LZ, The role of transporters in the pharmacokinetics of orally administered drugs, Pharmaceutical Research, 26(9), 2009, 2039-2054.
21. Clark DE, Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena of intestinal absorption, Journal of Pharmaceutical Sciences, 88(8), 1999, 807-814.
22. Winiwarter S, Ax F, LennernÃs H, Hallberg A, Pettersson C, KarlÃn A, Hydrogen bonding descriptors in the prediction of human *in vivo* intestinal permeability, Journal of Molecular Graphics and Modelling, 21(4), 2003, 273-287.
23. Alex A, Millan DS, Perez M, Wakenhuta F, Whitlocka GA, Intramolecular hydrogen bonding to improve membrane permeability and absorption in beyond rule of five chemical space, MedChemComm, 2, 2011, 669-674.
24. Kubinyi H, Nonlinear dependence of biological activity on hydrophobic character: the bilinear model, Farmaco-edizione scientifica, 34(3), 1979, 248-276.
25. Narayanaswamy R, Wai LK, Ismail IS, Molecular docking analysis of natural compounds as Human neutrophil elastase (HNE) inhibitors, Journal of Chemical and Pharmaceutical Research, 5(10), 2013, 337-341.
26. Bayat Z, Zanoozi M, DFT-Based QSAR Prediction of 1-Octanol/Water Partition Coefficient of Adamantine derivatives drugs, Journal of Chemical and Pharmaceutical Research, 2(6), 2010, 416-423.
27. Ertl P, Rohde B, Selzer P, Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties, Journal of Medicinal Chemistry, 43(20), 2000, 3714-3717.
28. Chen LL, Yao J, Yang JB, Yang J, Predicting MDCK cell permeation coefficients of organic molecules using membrane-interaction QSAR analysis, Acta Pharmacologica Sinica, 11, 2005, 1322-1333.
29. Billones LT, Billones JB, A univariate analysis of molecular properties and inhibitory activity of dihydrothiophenones against dihydroorotate dehydrogenase of malaria parasite, Journal of Chemical and Pharmaceutical Research, 6(8), 2014, 209-217.
30. Le TT, Hendriks AJ, Relationships between absorption efficiency of elements in mammals and chemical properties, Critical Reviews in Toxicology, 43(9), 2013, 800-809.
31. Hernandez MA, Rathinavelu A, Basic Pharmacology: Understanding Drug Actions and Reactions, CRC press publications, 2006, 33-34.
32. Kokate A, Li X, Jasti B, Effect of Drug Lipophilicity and Ionization on Permeability Across the Buccal Mucosa: A Technical Note, AAPS PharmSciTech, 9(2), 2008, 501-504.

Source of Support: Nil, **Conflict of Interest:** None.