Chem Biol Drug Des 2011; **78:** 112–123

Research Article

doi: 10.1111/j.1747-0285.2011.01106.x

Comparative QSAR Analyses of Competitive CYP2C9 Inhibitors using Three-Dimensional Molecular Descriptors

Viney Lather and Miguel X. Fernandes*

Centro de Química da Madeira, Departamento de Química, Universidade da Madeira, Campus da Penteada, 9000-390 Funchal, Portugal
**Corresponding author: Miguel X. Fernandes, mxf@uma.pt*
Present address: Viney Lather, Jan Nayak Ch. Devi Lal Memorial College of Pharmacy, Sirsa-125005, India.

One of the biggest challenges in QSAR studies using three-dimensional descriptors is to generate the bioactive conformation of the molecules. Comparative QSAR analyses have been performed on a dataset of 34 structurally diverse and competitive CYP2C9 inhibitors by generating their lowest energy conformers as well as additional multiple conformers for the calculation of molecular descriptors. Three-dimensional descriptors accounting for the spatial characteristics of the molecules calculated using E-Dragon were used as the independent variables. The robustness and the predictive performance of the developed models were verified using both the internal [leave-one-out (LOO)] and external statistical validation (test set of 12 inhibitors). The best models (MLR using GETAWAY descriptors and partial least squares using 3D-MoRSE) were obtained by using the multiple conformers for the calculation of descriptors and were selected based upon the higher external prediction (R^2_{test} values of 0.65 and 0.63, respectively) and lower root mean square error of prediction (0.48 and 0.48, respectively). The predictive ability of the best model, i.e., MLR using GETAWAY descriptors was additionally verified on an external test set of quinoline-4-carboxamide analogs and resulted in an R^2_{test} value of 0.6. These simple and alignment-independent QSAR models offer the possibility to predict CYP2C9 inhibitory activity of chemically diverse ligands in the absence of X-ray crystallographic information of target protein structure and can provide useful insights about the ADMET properties of candidate molecules in the early phases of drug discovery.

Key words: 3D-MoRSE, ADMET, CYP2C9, GETAWAY, QSAR, RDF, WHIM

Received 20 October 2009, revised 20 September 2010 and accepted for publication 13 February 2011

Cytochrome P450 (CYP) comprises a superfamily of hemoproteins that function as the terminal oxidase of the mixed function oxidase system and catalyze the metabolism of a large number of both exogenous and endogenous ligands in processes that can be beneficial or harmful for the organism (1). Of the approximately 57 human CYP genes cloned and classified according to sequence homology in families 1, 2, 3, 4, 5, 7, 8, 11, 17, 19, 21, 24, 27, and 51, three CYP families comprising of 1–3 and approximately 12 unique enzymes have been shown to play a substantial role in the human hepatic metabolism of drugs and non-drug xenobiotics (2–5). The remainder are of importance in the metabolism and/or biosynthesis of endogenous compounds, such as bile acids, biogenic amines, eicosanoids, fatty acids, phytoalexins, retinoids, and steroids (1). Although CYPs display high structural homology, they often have distinct roles in xenobiotic metabolism, with active sites that enable broad and overlapping substrate specificity, which is further complicated by ligand binding promiscuity.

CYP2C9, one of the four known members of the human CYP2C family, is one of the important drug-metabolizing CYP in humans (6) and is involved in the metabolism of commonly used polar acidic drugs (7). CYP2C9 is responsible for the metabolism of up to 15% of currently used therapeutics. CYP2C9 is the primary enzyme responsible for the metabolism of nonsteroidal anti-inflammatory drugs, oral antidiabetic agents, oral anticoagulants, and angiotensin-II receptor blockers (8). CYP2C9 is also the major enzyme involved in the disposition of warfarin. Some of the more potent CYP2C9 inhibitors include amiodarone, fluorouracil, metronidazole, miconazole (especially systemic use), and sulfamethoxazole (usually combined with trimethoprim) (6). All of the usual enzyme inducers, such as barbiturates, carbamazepine, and rifampin, can substantially increase CYP2C9 activity (6). The alteration of CYP2C9's activity plays a role in undesired side effects of drugs, especially those with low therapeutic indexes that are substrates of CYP2C9, and could produce severe consequences. The ability to predict, early in the drug development process, CYP2C9's inhibitory activity of lead compounds would therefore be extremely useful, especially to anticipate adverse results.

There are only two CYP2C9 structures cocrystallized with warfarin (9) and flurbiprofen (10), but there has been a steady progression in understanding the mechanism of CYP2C9 interaction with ligands using a number of approaches. These studies include descriptive structure activity relationship studies on tienilic acid derivatives

(11), phenytoin analogs, and bis-traizole antifungals to aid in understanding the substrate and inhibitor specificity of CYP2C9 (12), site-directed mutagenesis studies carried out to study the importance of the I-helix residues Ser286 and Asn289 for conferring specificity for substrates diclofenac and ibuprofen (13), and combined NMR and molecular modeling to assist in defining the positioning of substrates in CYP2C9 active site (14).

Regarding three-dimensional quantitative structure activity relationship (QSAR) studies of CYP2C9, Jones *et al.* (15) proposed a comparative molecular field analysis (CoMFA) model that described compounds that inhibited (S)-warfarin 7-hydroxylation and enabled LOO predictions of K_i. Ekins *et al.* (16) developed 3D-QSAR pharmacophore models using Catalyst by generating multiple conformations and compared results with 3D- and 4D-QSAR analyses using molecular surface-weighted holistic invariant molecular descriptors (MS WHIM) on a set of CYP2C9 inhibitors that inhibited tolbutamide and diclofenac 4′-hydroxylation (n = 9), on inhibitors that inhibited (S)-warfarin 7-hydroxylation (n = 29), and on inhibitors that inhibited tolbutamide 4-hydroxylation (n = 13) resulting in correlation coefficient (r) values of 0.91, 0.89, and 0.71, respectively. Afzelius *et al.* (17) studied the use of alignment-independent descriptors in ALMOND for obtaining the qualitative and quantitative predictions of the competitive inhibition of CYP2C9 on a series of structurally diverse compounds. The quantitative model generated by the partial least squares (PLS) analysis of GRIND descriptors using the experimental K_i values resulted in regression coefficient (r^2) value of 0.77 and cross-validated correlation coefficient (q^2) value of 0.60. The model was externally validated using 12 compounds and predicted 11 of 12 K_i values within 0.5 log units. In a subsequent report, Afzelius *et al.* (18) derived a conformer and alignment-independent 3D-QSAR model based on the flexible molecular interaction fields calculated in GRID and employed these fields and alignment-independent descriptors derived in ALMOND on a training set consisting of 22 diverse and flexible competitive inhibitors of CYP2C9. The model resulted in a R^2 of 0.81 and q^2 of 0.62. The predictive ability of the model was externally evaluated with a test set of 12 competitive inhibitors, and 11 were predicted within 0.5 log unit. No correlation coefficients for the test set were reported in these studies.

Because of the wide chemical diversity of CYP2C9 ligands, multiple sites of metabolism, and very limited availability of cocrystallized complexes, conformer selection still remains a big challenge in deriving quantitative models for CYP2C9 inhibition. Simple, versatile, and highly predictive QSAR models, overcoming the problem of lacking good protein structural information and simultaneously taking into account the ligand chemical diversity, would prove to be highly beneficial in the exploration of new chemical entities in the drug design and development process at very early stages. In this study, we carried out a comparative QSAR analysis on a dataset of 34 competitive, structurally diverse, and stereospecific CYP2C9 inhibitors earlier used by Afzelius *et al.* (18). Global lowest energy conformers of all the molecules in the dataset were generated and used for the calculation of 3D-descriptors. QSAR models were developed on a training set of 22 inhibitors using statistical techniques: PLS regression analysis and/or multiple linear regression (MLR). The predictability of these models had been evaluated internally using LOO cross-validation and externally using a test set of 12 compounds. However, one of the biggest challenges in QSAR studies using 3D-descriptors is to generate the bioactive conformation of the molecules. To account for this problem of bioactive conformation, these models were then compared with other QSAR models developed by generating multiple conformers of all the molecules in the dataset followed by PLS and MLR statistics, a methodology similar to the 4D-QSAR analysis, where multiple conformations are used to generate the predictive models. A type of alignment-independent 3D- and 4D-QSAR analyses by generating the lowest energy conformers and multiple conformers has been used for the prediction of CYP2C9 inhibitory activity. Molecular descriptors encoding the three-dimensional structural information included 3D Molecules Representation of Structure based on Electron Diffraction (3D-MoRSE), Geometry, Topology and Atom-Weights AssemblY (GETAWAY), Radial distribution Function (RDF), and weighted holistic invariant molecular (WHIM) descriptors and were used as the independent variable against the K_i; the dependent variable. This research work differs from the one carried out by Afzelius *et al.* (18) in terms of development of faster, simpler, and predictive models with 3D-descriptors without use of any commercial tool.

Methods and Material

Dataset for analysis

A dataset of 34 structurally diverse competitive inhibitors with K_i values, for CYP2C9 determined by diclofenac-4-hydroxylation, ranging from 0.28 to 245 μM has been used in the present study (18). Although the inhibitory capacity of CYP2C9 inhibitors has been extensively studied by several laboratories, the quality of the data used in modeling is crucial because there is great variability in kinetic constants for the same compounds between laboratories and, correspondingly, when different sources of enzyme such as recombinant CYP, human liver microsomes, hepatocytes, and liver slices are used. In the present study, we have used the data points reported earlier by Afzelius *et al.* (18). The chemical structures of the dataset molecules along with their experimental K_i values are shown in Table 1. The activity data have been converted to −logarithmic scale ($-\log_{10} K_i$) (Table 2) and subjected to QSAR analysis using 3D-descriptors. Tanimoto coefficients for similarity of the molecules in the training and test sets were calculated based on the daylight fingerprints taking Nicardipine, the most active molecule in the dataset for comparison, and values were found to range between 0.09 and 0.28, showing all the molecules were chemically diverse. The dataset was subdivided into a training set of 22 compounds and a test set of 12 different compounds as used earlier by Afzelius *et al.* (18).

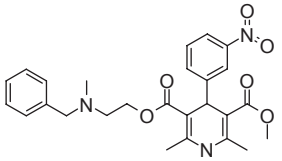
Computational details

Conformational analysis

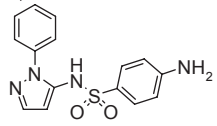
Each molecule in the dataset was encoded into a simplified molecular input line entry system (SMILES) string format (19). For stereospecific molecules (R/S), stereochemistry has been defined in the input molecules. Atomic 3D coordinates were generated by OMEGA

Table 1: Training set CYP2C9 inhibitors[a]

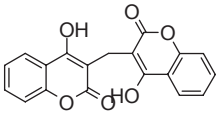
Nicardipine (**1**)
K_i = 0.28



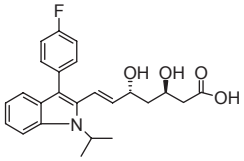
Sulphaphenazole (**2**)
K_i = 0.5



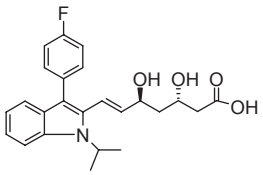
Dicoumarol (**3**)
K_i = 1.9



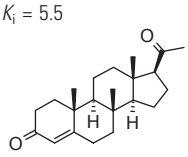
R, S-fluvastatin (**4**)
K_i = 2.2



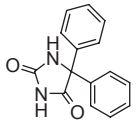
S, R-fluvastatin (**5**)
K_i = 3.3



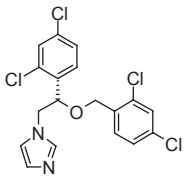
Progesterone (**6**)
K_i = 5.5



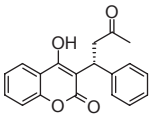
Phenytoin (**7**)
K_i = 6



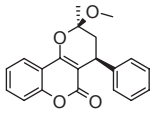
S-miconazole (**8**)
K_i = 6



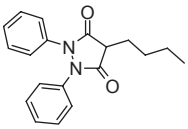
R-warfarin (**9**)
K_i = 13.6



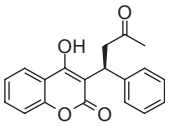
R, S-pyranocoumarin (**10**)
K_i = 16



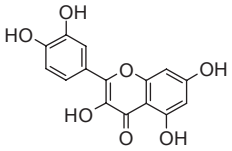
Phenylbutazone (**11**)
K_i = 19



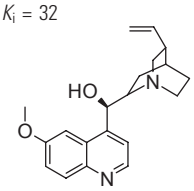
S-Warfarin (**12**)
K_i = 20



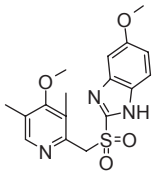
Quercetin (**13**)
K_i = 27



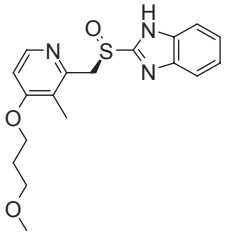
Quinine (**14**)
K_i = 32



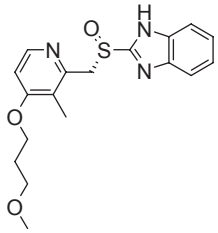
Omeprazole sulfone (**15**)
K_i = 35



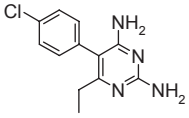
R-rabeprazole (**16**)
K_i = 36



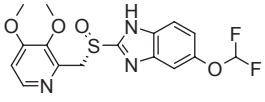
S-rabeprazole (**17**)
K_i = 37



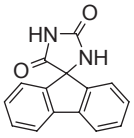
Pyrimethamine (**18**)
K_i = 51.5



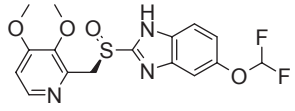
S-pantaprazole (**19**)
K_i = 64



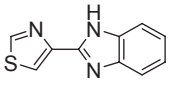
SFID (**20**)
K_i = 125



R-pantaprazole (**21**)
K_i = 145

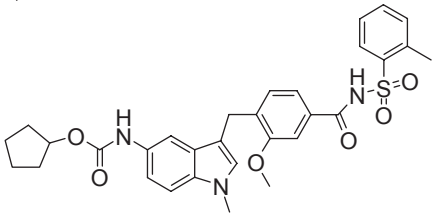


Thiabendazole (**22**)
K_i = 245

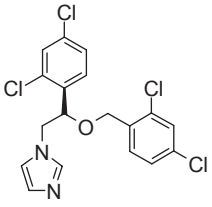


Test set CYP2C9 inhibitors[a]

Zafirlukast (**23**)
K_i = 2.5



R-miconazole (**24**)
K_i = 6



Kaempferol (**25**)
K_i = 6.0

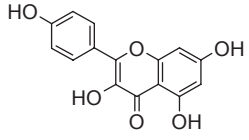
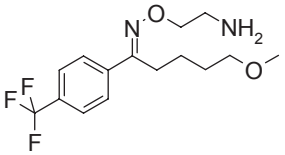
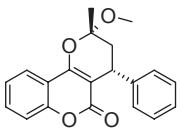


Table 1: *Continued*

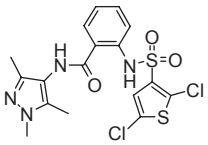
Fluvoxamine (**26**)
K_i = 8.5



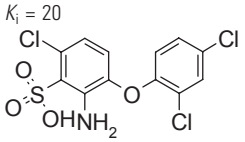
S, R-pyranocoumarin (**27**)
K_i = 16



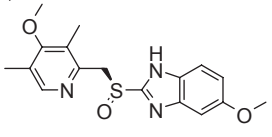
D-62126 (**28**)
K_i = 17.0



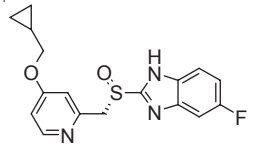
A-4438 (**29**)
K_i = 20



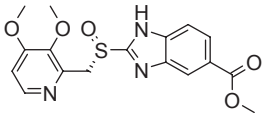
R-omeprazole (**30**)
K_i = 45



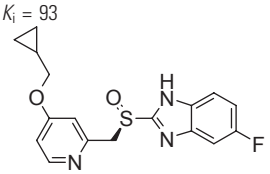
S-H259/31 (**31**)
K_i = 60



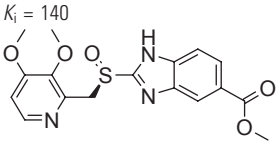
S-H287/23 (**32**)
K_i = 64



R-H259/31 (**33**)
K_i = 93



R-H287/23 (**34**)
K_i = 140



[a]Experimental values are given as micromolar.

version 2.2.1; Open Eye Scientific Software, Inc. (Santa Fe, NM, USA) OMEGA builds initial models of structures by assembling fragment templates along sigma bonds. Input molecule graphs are fragmented at exocyclic sigma and carbon to heteroatom acyclic (but not exocyclic) sigma bonds (20). Conformations for the fragments are either retrieved from pre-generated libraries built with *makefraglib* or constructed on-the-fly using the same distance constraints followed by geometry optimization protocol that *makefraglib* uses. Once an initial model is constructed, OMEGA generates additional models by enumerating ring conformations and invertible nitrogen atoms (Appendix S1). Ring conformations are taken from the same fragment library used to build an initial model.

For the generation of the first type of QSAR models (QSAR-I), each molecule was subjected to an energy minimization procedure using the molecular mechanics force field (MMFF) to generate the lowest energy conformation (Appendix S2). To generate the second type of models (QSAR-II), multiple conformers were generated for all the inhibitors by specifying the limit up to a maximum of 200 conformers/molecule in the input for OMEGA 2.0. The final geometries were obtained by subjecting all the conformations to energy refinement with semiempirical method AM1 using the MOPAC-7 program (Appendix S3). All geometries and electronic parameters were calculated in vacuum. The following sets of keywords were used in all quantum computations: AM1 PRECISE VECTORS BONDS PI KPOLAR ENPART.

Descriptors generation and selection

Conformational descriptors sensitive to the spatial positions of the atoms, whose values vary for the same molecule depending upon the selected conformer, were computed for the dataset molecules using E-DRAGON (21),[a] an electronic remote version of the descriptors calculating software DRAGON. The classes of descriptors calculated include 160 3D-MoRSE (22,23); 197 GETAWAY (24,25); 150 RDF (26,27); and 99 WHIM (28,29) descriptors. For the generation of QSAR-I models, lowest energy conformers finally optimized by AM1 semiempirical method were used and a total of 606 descriptors were calculated. The optimized structures were also used for the calculation of 1D- and 2D-descriptors using E-DRAGON. These descriptors include constitutional, molecular properties, topological, information theoretic indices, charge descriptors, topological charge indices, and edge adjacency indices. QSAR-I models were generated using these set of descriptors.

To generate the QSAR-II models, multiple conformations for each molecule were used for the generation of molecular descriptors. For each molecule and for each descriptor, the mean, the highest and lowest value, the range, and the standard deviation (SD) over the conformations were computed. This resulted in a total of 3030 descriptors for each molecule.

Selection of each class of descriptors was performed to reduce the pool of descriptors by eliminating those that satisfied at least one of the following conditions for the development of MLR models: (i) the descriptor has a zero/constant value for all the molecules investigated; (ii) the descriptors for the training set of molecules with a correlation coefficient <0.3 with the dependent variable (pK_i) were regarded as redundant; (iii) in the monoparametric correlation with (pK_i), the descriptor has a squared correlation coefficient lower than 0.1; (iv) in the monoparametric correlation, the descriptor has a *t*-test value lower than 0.1; (v) in the monoparametric correlation, the descriptor

Table 2: Experimental and calculated CYP2C9 inhibitory activity of training and test set molecules

S. No.	Experimental pK_i	PLS 3D-MoRSE QSAR-I	PLS GETAWAY QSAR-I	PLS 3D-MoRSE QSAR-II	MLR 3D-MoRSE QSAR-II	MLR GETAWAY QSAR-II
Training set						
1	6.55	6.55	6.20	6.34	6.51	6.35
2	6.30	5.66	5.45	5.58	5.71	5.20
3	5.72	5.45	5.60	5.20	5.41	5.50
4	5.66	5.29	5.60	5.56	5.61	5.68
5	5.48	5.54	5.69	5.59	5.55	5.75
6	5.26	5.14	5.34	5.02	5.08	5.47
7	5.22	5.16	4.86	5.11	4.82	5.44
8	5.22	5.29	5.22	5.50	5.32	5.14
9	4.87	5.09	4.99	5.24	5.11	5.00
10	4.80	5.34	5.36	5.20	5.00	4.88
11	4.72	5.05	4.91	5.13	5.30	4.94
12	4.70	5.10	4.89	5.20	4.88	4.74
13	4.57	4.34	4.62	4.45	4.53	5.06
14	4.49	4.50	4.66	4.59	4.42	4.37
15	4.46	4.41	4.77	4.37	4.53	4.68
16	4.44	4.40	4.42	4.38	4.99	4.34
17	4.43	4.17	4.53	4.27	4.23	4.08
18	4.29	4.33	4.33	4.40	3.95	4.42
19	4.19	3.85	3.63	3.76	3.88	4.27
20	3.90	4.52	4.38	4.33	4.06	3.96
21	3.84	3.83	3.62	3.74	3.93	3.72
22	3.61	3.70	3.64	3.75	3.89	3.74
Test set						
23	5.60	6.94	5.90	6.67	6.72	6.08
24	5.22	5.29	5.08	5.50	5.32	5.14
25	5.22	4.45	4.67	4.53	4.56	5.19
26	5.07	4.22	4.21	4.37	4.56	4.45
27	4.80	4.90	4.83	4.99	4.7	5.03
28	4.77	4.94	4.86	5.10	5.7	4.51
29	4.70	4.73	3.95	5.47	5.33	5.59
30	4.35	3.92	4.02	3.94	3.93	4.26
31	4.22	4.24	4.49	4.18	4.28	4.54
32	4.19	4.10	3.97	3.93	4.25	3.82
33	4.03	4.30	4.57	4.12	4.43	4.48
34	3.85	4.04	3.79	3.89	4.23	3.12

has a *F*-test value lower than 1 at a probability level of 0.05; (vi) highly correlated descriptors provide approximately identical information, if their pairwise correlation coefficient exceeded 0.75. Based on the intercorrelation coefficient values, one of the highly correlated descriptor was kept while others were removed.

Model development

As the predictability of a QSAR model is best judged by the external validation using a test set compounds, the dataset was divided into training and test sets (Appendix S4). We adopted multiple validation strategies like LOO cross-validation and external validation. As mentioned earlier, the kinetic data vary from one laboratory to another, and the dataset was divided into a training and test set of 22 and 12 compounds, respectively, and followed the criteria setup by Golbraikh *et al.* (30) (i) diversity of the training set, which is necessary condition for building a QSAR equation applicable to further compounds of interest in the same chemical domain; (ii) closeness of the representative points of both the training and test set in the descriptor space that ensures a proper validation of the model. Also, all the data points in the test set should fall in between the max and min activity values (range) for the training set of compounds.

QSAR models were generated separately for each class of descriptors using pK_i values for CYP2C9 inhibition as the dependent variable. For the development of QSAR models, the statistical techniques used were PLS and stepwise MLR. Stepwise MLR was used to study the influence of most important descriptors in the prediction. Standardization of variables, stepwise MLR, and PLS were performed using the statistical software STATISTICA version 7.0.[b]

Statistical methods

Multivariate methods establish relationships between predictor (independent) variables, *X*, and response (dependent) variables, *Y*, extracting factors from Y^TY and X^TX matrices. In partial least squares regression, PLS, the factors are extracted from Y^TXX^TY matrix, which is less restrictive and can be applied to situations where other multivariate methods fail (31). For instance, it can handle data with strongly correlated and/or noisy or numerous independent *X* variables, and a regression model from PLS can be

expected to have a smaller number of components without an appreciably smaller R^2 value.

PLS was applied to correlate each class of descriptors separately with the observed pK_i values. Because the variance associated with different descriptors can be very different, descriptors were autoscaled so as to assign unit variance to each descriptor. The optimum number of components in each PLS model generated was determined using the following criteria: (i) squared correlation coefficient; R^2 and (ii) LOO cross-validation; q^2.

Stepwise MLR is a model-building technique that finds subsets of predictor variables that most adequately forecast responses on a dependent variable by linear regression, given the specified criteria for adequacy of model fit (32). The basic procedure involves: (i) identifying an initial model; (ii) iteratively stepping or repeatedly altering the model at the previous step by adding or removing a predictor variable in accordance with the 'stepping criteria' ($F = 1$ for inclusion; $F < 1$ for exclusion); and (iii) terminating the search when stepping is no longer possible given the stepping criteria, or when a specified maximum number of steps has been reached. Specifically at each step, all the variables are reviewed and evaluated to determine which one will contribute the most to the equation. That variable is then included in the model, and the process starts again.

The statistical quality of the models was checked by parameters, such as squared correlation coefficient (R^2), adjusted R^2 (R^2_a), and variance ratio (F) at specified degrees of freedom (df). Cross-validated correlation coefficient (q^2) is calculated according to the formula:

$$q^2 = 1 - \frac{\sum (Y_{obs} - Y_{pred})^2}{\sum (Y_{obs} - \hat{Y})^2} \quad (1)$$

In the above equation, $\hat{Y}$ means average activity value of the training set, whereas Y_{obs} and Y_{pred} represent observed and LOO-predicted activity values of the training set. PRESS is given by the expression:

$$\text{PRESS} = \sum (Y_{obs} - Y_{pred})^2 \quad (2)$$

We verified the requirements formulated by Golbraikh and Trophsa (33) for considering a QSAR model as highly predictive if they satisfy the following conditions: (i) $q^2 > 0.5$; (ii) $R^2_{test} > 0.5$ (R^2_{test} is the correlation coefficient for test set predictions); (iii) R^2_0 or R'^2_0 should be close to R^2_{test} such that [value of $(R^2_{test} - R^2_0)/R^2_{test}$) or $(R^2_{test} - R'^2_0)/R^2_{test}$) is <0.1; ($R^2_0$ and R'^2_0 are correlation coefficients for regressions through the origin for predicted versus observed activities and for observed versus predicted activities, respectively], and $0.85 \le k \le 1.15$ or $0.85 \le K' \le 1.15$ (k and K' are the corresponding slopes of regression lines through the origin).

Roy and Roy (34) previously showed that the use of R^2_{test} might not be sufficient to indicate the external validation characteristics; R^2_m was used to be a measure of external prediction and was calculated as

$$R^2_m = R^2 \left(1 - \sqrt{|R^2 - R^2_0|}\right) \quad (3)$$

A value of $R^2_m > 0.5$ may be taken as an indicator of good external predictability.

Results and discussion

QSAR-I models

Partial least squares

The PLS regression was carried out using each class of 3D-descriptors as described earlier, generated for the lowest energy conformations, after removing the variables with zero/constant values and the variables with smaller coefficients as described above, until no further improvement was seen in q^2 value irrespective of the number of components. To avoid overfitting, the significance of each consecutive PLS component is examined, and it is stopped when the components are non-significant, i.e., no further improvement was seen in q^2 values. The statistical details of PLS models generated with each class of descriptors are shown in Table 3.

The PLS 3D-MoRSE model could explain 77% of the variance (adjusted coefficient of variation). The optimal number of latent variables for this PLS model was 3. The PLS 3D-MoRSE model for lowest energy conformers generated a LOO q^2 value of 0.50. Simple squared correlation coefficient R^2_{test} between the observed and predicted values of the test set compounds was found to be 0.55. Setting the intercept to zero, the squared correlation coefficient was found to be 0.54. As R^2 and R^2_0 values were not much different, an acceptable value of R^2_m (0.51) was obtained, which validated the predictability of the PLS 3D-MoRSE model.

Table 3: Comparative table of statistical analyses of different PLS QSAR-I models

Model	R^2	R^2_{adj}	q^2	R^2_{test}	R^2_0	K	R^2_m	RMSEP	PRESS
PLS 3D-MoRSE	**0.84**	**0.77**	**0.50**	**0.55**	**0.54**	**1.00**	**0.51**	**0.54**	**6.18**
PLS GETAWAY	**0.83**	**0.75**	**0.50**	**0.53**	**0.50**	**1.00**	**0.50**	**0.43**	**6.59**
PLS RDF	0.43	0.28	0.36	0.55	0.54	1.02	0.50	0.52	7.91
PLS WHIM	0.70	0.60	0.25	0.59	0.58	1.04	0.53	0.45	10.23
PLS 1D- and 2D-descriptors from E-DRAGON	0.64	0.56	0.20	0.47	0.46	1.02	0.42	0.65	13.68

Bold represents significant models.

The PLS GETAWAY model could explain 75% of the variance (R^2_{adj}). The optimal number of PLS components for this model was 3. The PLS GETAWAY model for lowest energy conformers generated a LOO q^2 value of 0.50. R^2_{test} between the observed and predicted values of the test set compounds was found to be 0.52. Setting the intercept to zero, the squared correlation coefficient was found to be 0.50. R^2_m was found to be 0.50, which validates the predictability of this model.

The other models were generated using RDF and WHIM 3D-descriptors. However, these models were able to explain a variance of only 30% and 60%, respectively. The numbers of PLS components used in these models were 3 and 4, respectively. The comparative PLS results using these classes of descriptors to that of the 3D-MoRSE and GETAWAY are shown in Table 3. The PLS RDF and PLS WHIM models were of lower significance in explaining the LOO variance as shown by low q^2 values and high PRESS statistics.

Forward stepwise MLR

As mentioned earlier, the initial pool of each class of 3D-descriptors was reduced by eliminating the redundant variables followed by subjection to MLR using forward feature selection criteria.

The MLR model using the 3D-MoRSE descriptors resulted in a three-parametric equation:

pK_i = −0.393 (±0.087) Mor05m + 1.039 (±0.219) Mor16u − 0.455 (±0.164) Mor08m + 1.692 (±0.368)

$n_{training}$ = 22, R^2 = 0.83, R^2_{adj} = 0.81, F = 30.01 (df 3, 18), q^2 = 0.76, SEE = 0.34, p < 0.000

PRESS = 2.94, n_{test} = 12, R^2_{test} = 0.30, R^2_0 = 0.297, R^2_m = 0.28

This trivariant model was able to explain 81% of the variance (R^2_{adj}), and LOO-predicted variance was found to be 76.3%. However, the predictability of this model for the test set compounds was found to be very low.

The MLR model using the GETAWAY descriptors also resulted in a following three-parametric equation:

pK_i = 0.786 (±0.070) RTv − 1.778 (±0.240) H2u − 7.997 (±1.484) R4m+ + 2.947 (±0.379)

$n_{training}$ = 22, R^2 = 0.89, R^2_{adj} = 0.87, F = 48.81 (df 3, 18), q^2 = 0.81, SEE = 0.27, p < 0.000

PRESS = 2.35, n_{test} = 12, R^2_{test} = 0.26, R^2_0 = 0.26, R^2_m = 0.26

The MLR GETAWAY model was able to explain 87% of the variance, and LOO-predicted variance was 81%. The predictability of this model for the test set compounds was found to be low.

The MLR models were also developed with RDF and WHIM 3D-descriptors. These models were able to explain only 66% and 45% of the variance, respectively. The comparative results obtained with these descriptors to that of 3D-MoRSE and GETAWAY are shown in Table 4. As it can be seen, there are remarkable differences concerning the explanation of the experimental variance given by these models compared to that of 3D-MoRSE and GETAWAY descriptors. The meanings of the 3D-variables used in the various MLR models developed in the current work are defined in Table 5.

The advantage of obtaining statistically significant QSAR models using 3D-descriptors was established by the development of QSAR models using 1D- and 2D-descriptors. The results are shown in Tables 3 and 4. None of the models using either PLS or MLR resulted in a statistically significant model.

QSAR-II models

QSAR-II models were developed by using the multiple conformers for each molecule. The mean, the highest and lowest value, the range, and the S. D. values for each descriptor were computed for all the molecules. The calculated descriptors were subjected to PLS and stepwise MLR analyses. Separate models were generated for each class of 3D-descriptors. For each class of descriptors, different combinations of the subclasses, i.e., mean, highest and lowest values, range, and the SD values were used to generate the suitable models. Only those analyses, which resulted in statistically valid models, are described later.

Partial least squares

The PLS regression was performed with each subclass and combination thereof for the 3D-descriptors generated for the multiple conformers, after removing the variables with smaller coefficients, until no further improvement was seen in q^2 value irrespective of the number of components. The statistical details of the PLS models generated with each class of descriptors are shown in Table 6.

Table 4: Comparative table of statistical analyses of different MLR QSAR-I models

Model	R^2	R^2_{adj}	q^2	R^2_{test}	R^2_0	K	R^2_m	RMSEP	PRESS
MLR 3D-MoRSE	0.83	0.81	0.76	0.30	0.29	1.03	0.28	0.75	2.94
MLR GETAWAY	0.89	0.87	0.81	0.26	0.25	0.92	0.24	0.78	2.35
MLR RDF	0.71	0.66	0.58	0.20	0.20	1.07	0.20	1.03	5.62
MLR WHIM	0.53	0.45	0.22	0.02	–	–	–	–	–
MLR 1D- and 2D-descriptors from E-DRAGON	0.88	0.85	0.78	0.27	0.17	–	–	–	–

Table 5: 3D-descriptors of the MLR QSAR models reported in this study

Descriptor	Class	Definition
QSAR-I models		
Mor05m	3D-MoRSE	3D-MoRSE – signal 05/weighted by atomic masses
Mor16u		3D-MoRSE – signal 16/unweighted
Mor08m		3D-MoRSE – signal 08/weighted by atomic masses
RTv	GETAWAY	*R* total index/weighted by atomic van der Waals volumes
H2u		*H* autocorrelation of lag 2/unweighted
R4m+		*R* maximal autocorrelation of lag 4/weighted by atomic masses
RDF060m	RDF	Radial Distribution Function –6.0/weighted by atomic masses
RDF030m		Radial Distribution Function –3.0/weighted by atomic masses
RDF075p		Radial Distribution Function –7.5/weighted by atomic polarizabilities
E2m	WHIM	2nd component accessibility directional WHIM index/weighted by atomic masses
E3e		3rd component accessibility directional WHIM index/weighted by atomic Sanderson electronegativities
G1s		1st component symmetry directional WHIM index/weighted by atomic electrotopological states
QSAR-II models		
Mor05m	3D-MoRSE	3D-MoRSE – signal 05/weighted by atomic masses
Mor16e		3D-MoRSE – signal 16/weighted by atomic Sanderson electronegativities
Mor22e		3D-MoRSE – signal 22/weighted by atomic Sanderson electronegativities
R2u+	GETAWAY	*R* maximal autocorrelation of lag 2/unweighted
R8e+		*R* maximal autocorrelation of lag 8/weighted by atomic Sanderson electronegativities
R5m+		*R* maximal autocorrelation of lag 4/weighted by atomic masses
RDF050v	RDF	Radial Distribution Function –6.0/weighted by atomic van der Waals volumes
RDF020u		Radial Distribution Function –2.0/unweighted
RDF020m		Radial Distribution Function –2.0/weighted by atomic masses
L3e	WHIM	3rd component size directional WHIM index/weighted by atomic Sanderson electronegativities
E2e		2nd component accessibility directional WHIM index/weighted by atomic Sanderson electronegativities

Table 6: Comparative table of statistical analyses of different PLS QSAR-II models

Model	R^2	R^2_{adj}	q^2	R^2_{test}	R^2_0	K	R^2_m	RMSEP	PRESS
PLS 3D-MoRSE	**0.82**	**0.74**	**0.51**	**0.63**	**0.61**	**1.01**	**0.54**	**0.48**	**5.94**
PLS GETAWAY	0.77	0.68	0.43	0.41	0.34	1.00	0.31	0.48	7.16
PLS RDF	0.66	0.55	0.27	0.49	0.41	1.02	0.35	0.41	7.16
PLS WHIM	**0.74**	**0.64**	**0.50**	**0.59**	**0.58**	**1.05**	**0.53**	**0.49**	**6.20**

The best model was obtained by using the average values of 3D-MoRSE descriptors and could explain 74% of the experimental variance (adjusted coefficient of variation). The number of latent variables for the PLS equation was found to be 3. The LOO-predicted variance was found to be 51%. The R^2_{test} value for the test set molecules was found to be 0.63. Setting intercept to zero, the squared correlation coefficient was found to be 0.61. An acceptable value of R^2_m (0.54) was obtained, indicating the good predictability of the model. Figure 1 shows the fit plots of the predicted versus experimental pk_i values for CYP2C9 inhibition of the training and test sets derived from the PLS 3D-MoRSE QSAR-II model.

PLS models using GETAWAY (three PLS components) and RDF (three PLS components) descriptors resulted in explaining 68% and 55% of the experimental variance, respectively. The LOO-predicted variance using these descriptors was found to be 43% and 27%, respectively, explaining the poor statistical validation.

PLS regression model developed using WHIM descriptors (five PLS components) was able to explain 64% of experimental variance. The LOO-cross-validated predicted variance and the external predictability for the model using WHIM descriptors were found to be 50% and 59%, respectively. An acceptable value of R^2_m (0.54) was obtained, indicating the good predictability of the model. The PRESS and RMSEP statistics were found to be 6.20 and 0.49.

Stepwise MLR

MLR 3D-MoRSE approach. A total of 800 3D-MoRSE descriptors were computed and subjected to reduction by eliminating the redundant variables followed by forward stepwise MLR. The average values of the descriptors calculated from the multiple conformers of each molecule resulted in the best four-parametric model. This model is shown below:

pK_i = –0.592 (±0.078) Mor05m + 1.099 (±0.223) Mor16e + 0.521 (±0.161) Mor22e + 1.715 (±0.621) + 2.280 (±0.356)

$n_{training}$ = 22, R^2 = 0.85, R^2_{adj} = 0.82, F = 24.60 (df 4, 17), q^2 = 0.76, SEE = 0.32, p < 0.0000

PRESS = 2.96, n_{test} = 12, R^2_{test} = 0.51, R^2_0 = 0.51, R^2_m = 0.51, RMSEP = 0.56

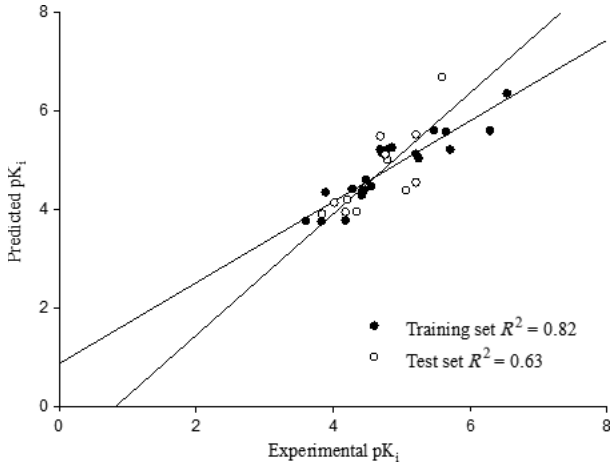


Figure 1: Plots of predicted versus experimental pK_i of the training and test set molecules based on the PLS 3D-MoRSE QSAR-II Model.

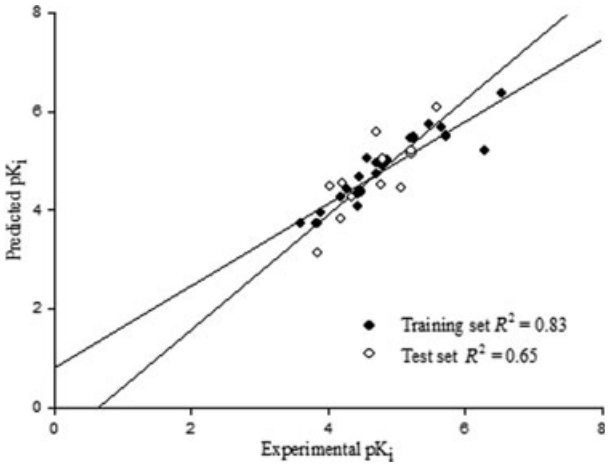


Figure 2: Plots of predicted versus experimental pK_i of the training and test set molecules based on the MLR GETAWAY QSAR-II Model.

This model involving four descriptors could explain 82% of the variance (R^2_{adj}), and LOO-predicted variance was found to be 76%. The Fisher value was found to be 24.60 (on 4 and 17 df). The values of external prediction parameters R^2_{test}, R^2_0, and R^2_m were found to be 0.51, 0.51, and 0.51, respectively. These parameters validated the predictability of this model. The descriptors involved in this model are defined in Table 5.

The other models were also developed using each of the subclasses of descriptors; however, no suitable models were generated showing the importance of mean values of 3D-MoRSE descriptors calculated for multiple conformers of each molecule in the CYP2C9 dataset.

MLR GETAWAY approach. Initial pool of each subclass of GETAWAY descriptors was reduced by eliminating the redundant variables followed by MLR using forward stepwise feature selection. The most significant model using the GETAWAY descriptors was derived from the average values derived from the multiple conformers of each molecule. Setting the 'stepping criteria' ($F = 1$ for inclusion; $F < 1$ for exclusion), the following equation was obtained:

pK_i = −75.132 (±8.936) R2u+ + 69.996 (±21.708) R8e+ − 5.409 (±1.914) R5m+ + 8.755 (±0.541)

$n_{training}$ = 22, R^2 = 0.81, R^2_{adj} = 0.78, F = 25.16 (df 3, 18), q^2 = 0.70, SEE = 0.36, p < 0.0000

PRESS = 3.01, n_{test} = 12, R^2_{test} = 0.65, R^2_0 = 0.63, R^2_m = 0.56, RMSEP = 0.48

The standard errors of regression coefficients are given within parenthesis. This trivariant model could explain 78% of the experimental variance (R^2_{adj}), and LOO-predicted variance was found to be 70%. The F statistic (on 3 and 18 df) for this model was found to be 25.16 with a p-value of <0.000. All the t-values were significant with low p-values, which confirmed the significance of each descriptor. The predictive ability of the model was validated on the test set resulting in R^2_{test} value of 0.65. R^2_m value was found to be 0.55, which further validated the external predictability of this model. The PRESS statistic was found to be 3.91. Figure 2 shows the fit plots of the predicted versus experimental pk_i values for CYP2C9 inhibition of the training and test sets derived from the MLR GETAWAY QSAR-II model.

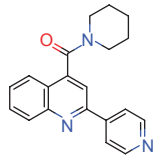
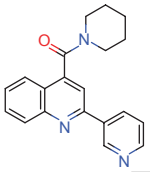
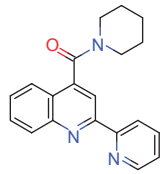
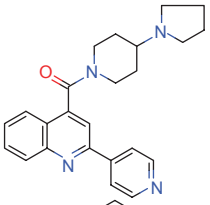
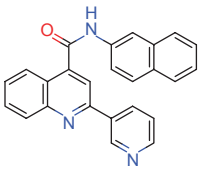
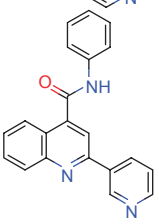
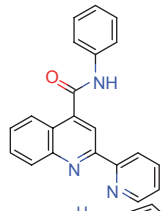
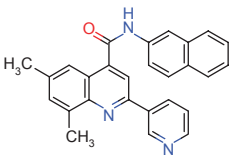
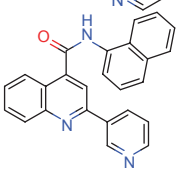
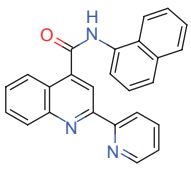
GETAWAY descriptors are calculated from the leverage matrix obtained by the centered atomic coordinates (molecular influence matrix, MIM). GETAWAY descriptors are geometrical descriptors encoding information on the effective position of substituents and fragments in the molecular space. These descriptors are independent of molecular alignment and account for information on molecular size and shape as well as for specific atomic properties. R and R+ descriptors are obtained from the leverage/geometric matrix (24,25).

The descriptors involved in the MLR GETAWAY model are belonging to R subcategory of GETAWAY descriptors; it is clear that CYP2C9

Table 7: Comparative table of statistical analyses of different MLR QSAR-II models

Model	R^2	R^2_{adj}	q^2	R^2_{test}	R^2_0	K	R^2_m	RMSEP	PRESS
MLR 3D-MoRSE	**0.85**	**0.82**	**0.76**	**0.51**	**0.51**	**1.04**	**0.51**	**0.56**	**2.96**
MLR GETAWAY	**0.83**	**0.81**	**0.70**	**0.65**	**0.63**	**1.01**	**0.54**	**0.48**	**3.01**
MLR RDF	0.77	0.74	0.57	0.10	–	–	–	0.72	5.88
MLR WHIM	0.66	0.62	0.58	0.31	0.11	1.02	0.17	0.50	5.33

Table 8: Experimental and calculated CYP2C9 inhibitory activity of Quinoline-4-carboxamide analogs by MLR using GETAWAY descriptors QSAR model

Compound	Structure	Exp. pK_i	Pred. pK_i	Compound	Structure	Exp. pK_i	Pred. pK_i
1		4.42	5.00	**2**		4.14	5.01
3		4.28	4.63	**4**		5.15	5.44
5		5.11	5.94	**6**		4.68	5.39
7		4.61	5.46	**8**		5.06	5.07
9		5.16	5.90	**10**		5.16	5.88

inhibitory activity is influenced by the molecular size and shape of the molecules in the dataset.

MLR RDF and MLR WHIM approaches. Forward stepwise MLR models were generated using each subclass and the combination thereof for RDF and WHIM descriptors, respectively. However, no statistically valid models were generated with these classes of descriptors. Table 7 shows the comparative results of these classes of descriptors to that of 3D-MoRSE and GETAWAY descriptors. The best model using RDF descriptors resulted in a three-parametric equation and had an *F*-test value of 20.47 (on 3 and 18 df). The MLR model using WHIM descriptors resulted in a two-parameter equation and could explain 62% of variance in the activity. However, the predictability of the models derived from RDF and WHIM descriptors was found to be very low, as shown in Table 7 that the MLR models based on these descriptors had R^2_{test} values of 0.10 and 0.31, respectively.

Validation of best QSAR model

Further validation of best QSAR model, i.e., MLR using GETAWAY descriptors was carried out on a test set of ten quinoline-4-carboxamide analogs (35). The CYP2C9 inhibitory activity (K_i) of these molecules was also determined using dicolfenac-4-hydroxylation by Peng *et al.* (35). Multiple conformers of all the structures in external dataset were generated by specifying the limit up to a maximum of 200 conformers/molecules. The final geometries were obtained by subjecting all the conformers to energy refinement with semiempirical method AM1. GETAWAY descriptors were calculated by using E-DRAGON.

The predictability of the MLR model using GETAWAY descriptors resulted in an R^2_{test} value of 0.6 for quinoline-4-carboxamide analogs, which further validates this present mathematical modeling approach for the prediction of CYP2C9 inhibitory activity of chemically diverse molecules. The results are produced in Table 8.

Overview and Conclusion

PLS and stepwise MLR approaches have been applied for the linear modeling of chemically diverse CYP2C9 inhibitors using 3D-descriptors (3D-MoRSE, GETAWAY, RDF, WHIM). Because bioactive

conformer generation and selection remain a major challenge in using the 3D-descriptors for the development of QSAR models, the current work employs the use of lowest energy conformers for generating the QSAR models and then comparing these with that of the models developed by using multiple conformers for all the molecules in the CYP2C9 dataset. The predictive ability of the models was estimated from the prediction of the CYP2C9 inhibitory activity of the test set of 12 compounds. Comparative results of the various PLS and MLR models developed are shown in Tables 3, 4, 6 and 7. The best models were found to be those based on the use of multiple conformers compared to that of their lower energy conformer counterpart. The best QSAR model was obtained by stepwise MLR using GETAWAY descriptors calculated using the multiple conformers and resulted in a high external prediction (R^2_{test} = 0.64) with low RMSEP value (0.48). The second most statistically significant model was based on PLS regression using the 3D-MoRSE descriptors based on the internal (q^2 = 0.51) and external (R^2_{test} = 0.63) predictive power with the low RMSEP value (0.48). Further, based upon R^2_m values, which accounts for the large differences between observed and predicted values, the MLR GETAWAY and PLS 3D-MoRSE model derived by using multiple conformers were found to be superior (R^2_m = 0.54) in comparison with other models developed. The other statistically valid models were derived from PLS regression of WHIM descriptors (q^2 = 0.50, R^2_{test} = 0.59) and stepwise MLR analysis of 3D-MoRSE descriptors generated using the multiple conformers. The models developed using the multiple conformers were more predictive when compared to that of the models developed by using the lower energy conformers. The only statistically significant models using the lower energy conformers were derived by PLS regression of 3D-MoRSE (q^2 = 0.50, R^2_{test} = 0.55) and GETAWAY descriptors (q^2 = 0.50, R^2_{test} = 0.52). None of the models derived based on stepwise MLR analysis resulted in a statistically valid prediction.

This study differs from the one carried out by Afzelius *et al.* in that these QSAR models are simple with good external prediction and were developed in a conformational-dependent as well as the independent manner using the lowest energy conformers. This methodology could prove of immense help where the bioactive conformations of the ligands are unknown owing to their lack of bound X-ray or NMR crystallographic information for the CYP2C9 enzyme. The 3D-descriptors used in the study are easy to interpret and calculate.

Overall, a conformational independent and dependent methodology similar to the 3D- and 4D-QSAR approaches have been used resulting in simple, versatile, and fast predictive models for the prediction of CYP2C9 inhibitory activity of chemically diverse inhibitors. These simple QSAR models for the prediction of CYP2C9 inhibitory activity could prove very useful in exploring the ADMET fate of new chemical moieties in the early stages of drug design and discovery.

Acknowledgments

We thank Fundação para a Ciência e Tecnologia (Portugal) for Grant SFRH/BPD/30954/2006 attributed to Viney Lather.

References

1. Nelson D.R., Koymans L., Kamataki T., Steqeman J.J., Feyereisen R., Waxman D.J., Waterman M.R., Gotoh O., Coon M.J., Estabrook R.W., Gunsalus I.C., Nebert D.W. (1996) P450 superfamily: update on new sequences, gene mapping, accession numbers and nomenclature. Pharmacogenetics;6:1–42.
2. Wrighton S.A., Stevens J.C. (1992) The human hepatic cytochromes P450 involved in drug metabolism. Crit Rev Toxicol;22:1–21.
3. Gonzalez F.J. (1992) Human cytochromes P450: problems and prospects. Trends Pharmacol Sci;13:346–352.
4. Nelson D.R., Zeldin D.C., Hoffman S M.G., Maltais L.J., Wain H.M., Nebert D.W. (2004) Comparison of cytochrome P450 (CYP) genes from the mouse and human genomes, including nomenclature recommendations for genes, pseudogenes and alternative-splice variants. Pharmacogenetics;14:1–18.
5. Cholerton S., Daly A.K., Idle J.R. (1992) The role of individual human cytochromes P450 in drug metabolism and clinical response. Trends Pharmcol Sci;13:434–439.
6. Miners J.O., Birkett D.J. (1998) Cytochrome P4502C9: an enzyme of major importance in human drug metabolism. Br J Clin Pharmacol;45:525–538.
7. Hall S.D., Hamman M.A., Rettie A.E., Wienkers L.C., Trager W.F., VandenBranden M., Wrighton S.A. (1994) Relationships between the levels of cytochrome P4502C9 and its prototypic catalytic activities in human liver microsomes. Drug Metab Dispos;22:975–977.
8. Rettie A.E., Jones J.P. (2005) Clinical and toxicological relevance of CYP2C9: drug-drug interactions and pharmacogenetics. Ann Rev Pharmacol Toxicol;45:477–494.
9. Williams P.A., Cosme J., Ward A., Angove H.C., Matak Vinkovic D., Jhoti H. (2003) Crystal structure of human cytochrome P450 2C9 with bound warfarin. Nature;424:464–468.
10. Wester M.R., Yano J.K., Schoch G.A., Yang C., Griffin K.J., Stout C.D., Johnson E.F. (2004) The structure of human cytochrome P450 2C9 complexed with flurbiprofen at 2.0 A resolution. J Biol Chem;279:35630–35637.
11. Mancy A., Broto P., Dijols S., Dansette P.M., Mansuy D. (1995) The substrate binding site of human liver cytochrome P4502C9: an approach using designed tienilic acid derivatives and molecular modeling. Biochemistry;34:10365–10375.
12. Morsman J.M., Smith D.A., Jones B.C., Hawksworth G.M. (1995) Role of hydrogen-bonding in substrate structure-activity relationships for CYP2C9. ISSX Proc;8:259.
13. Klose T.S., Ibeanu G.C., Ghanayem B.I., Pedersen L.G., Li L., Hall S.D., Goldstein J.A. (1998) Identification of residues 286 and 289 as critical for conferring substrate specificity of human CYP2C9 for diclofenac and ibuprofen. Arch Biochem Biophys;357:240–248.
14. Poli-Scaife S., Attias R., Dansette P.M., Mansuy D. (1997) The substrate binding site of human liver cytochrome P4502C9: an NMR study. Biochemistry;36:12672–12682.
15. Jones J.P., He M., Trager W.F., Rettie A.E. (1996b) Three-dimensional quantitative structure activity relationship for inhibitors of cytochrome P4502C9. Drug Metab Dispos;24:1–6.

16. Ekins S., Bravi G., Binkley S., Gillespie J.S., Ring B.J., Wikel J.H., Wrighton S.A. (2000) Three- and four-dimensional quantitative structure activity relationship (3D/4D-QSAR) analyses of CYP2C9 inhibitors. Drug Metab Dispos;28:994–1002.
17. Afzelius L., Masimirembwa C.M., Karlen A., Andersson T.B., Zamora I. (2002) Discriminant and quantitative PLS analysis of competitive CYP2C9 inhibitors versus non-inhibitors using alignment independent GRIND descriptors. J Comp Aid Mol Des;16:443–458.
18. Afzelius L., Zamora I., Masimirembwa C.M., Karlen A., Andersson T.B., Mecucci S., Baroni M., Cruciani G. (2004) Conformer- and alignment-independent model for predicting structurally diverse competitive CYP2C9 inhibitors. J Med Chem;47:907–914.
19. Weininger D. (1998) SMILES 1. Introduction and encoding rules. J Chem Inf Comput Sci;28:31.
20. Bostrom J., Greenwood J.R., Gottfries J. (2003) Assessing the performance of OMEGA with respect to retrieving bioactive conformations. J Mol Graph Model;21:449–462.
21. Tetko I.V., Gasteiger J., Todeschini R., Mauri A., Livingstone D., Ertl P., Palyulin V.A., Rodchenko E.V., Zefirov N.S., Makarenko A.S., Tanchuk V.Y., Prokopenko V.V. (2005) Virtual computational chemistry laboratory-design and description. J Comput Aid Mol Des;19:453–463.
22. Schuur J.H., Seizer P., Gasteiger J. (1996) The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure-spectra correlations and studies of biological activity. J Chem Inf Comput Sci;36:334–344.
23. Saiz-Urra L., Gonzalez M.P., Teijeira M. (2006) QSAR studies about cytotoxicity of benzophenazines with dual inhibition toward both topoisomerases I and II: 3D-MoRSE descriptors and statistical considerations about variable selection. Bioorg Med Chem;14:7347–7358.
24. Consonni V., Todeschini R., Pavan M. (2002) Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors. J Chem Inf Comput Sci;42:682–692.
25. Gonzalez M.P., Teran C., Teijeira M., Gonzalez-Moa M.J. (2005) GETAWAY descriptors to predicting A_{2A} adenosine receptor agonists. Eur J Med Chem;40:1080–1086.
26. Todeschini R., Consonni V. (2000) Handbook of Molecular Descriptors. Weinheim, Germany: Wiley-VCH.
27. Gonzalez M.P., Teran C., Teijeira M., Helguera A.M. (2006) Radial distribution function descriptors: an alternative for predicting A2A adenosine receptor agonists. Eur J Med Chem;41:56–62.
28. Bravi G., Gancia E., Mascani P., Pegna M., Todeschini R., Zaliani A. (1997) MS-WHIM, new 3D theoretical descriptors derived from molecular surface properties: a comparative 3D-QSAR study in a series of steroids. J Comput Aid Mol Des;11:79–92.
29. Bravi G., Wikel J.H. (2000a) Application of MS-WHIM descriptors 1. Introduction of new molecular surface properties and 2. Prediction of binding affinity data. Quant Struct Act Rel;19:29–38.
30. Golbraikh A., Trophsa A. (2000) Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. Mol Div;5:231–234.
31. Wold S., Eriksson L. (1995) Validation tools. In: van de Waterbeemd H., editor. Chemometric Methods in Molecular Design. Weinheim: VCH; p. 312–317.
32. Darlington R.B. (1990) Regression and Linear Models. New York: McGraw-Hill.
33. Golbraikh A., Trophsa A. (2002) Beware of q2! J Mol Graph Model;20:269–276.
34. Roy P., Roy K. (2008) On some aspects of variable selection for partial least squares regression models. QSAR Comb Sci;27:302–313.
35. Peng C., Cape J.L., Rushmore T., Crouch G.J., Jones J.P. (2008) Cytochrome P450 2C9 Type II binding studies on Quinoline-4-carboxamide analogs. J Med Chem;51:8000–8011.

Notes

[a]VCCLAB, Virtual Computational Chemistry Laboratory, available at: http://www.vcclab.org.

[b]STATISTICA (data analysis software system), version 7, Statsoft Inc., available at: http://www.statsoft.com.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix S1. CYP2C9_Multiconfs_QSARII.

Appendix S2. CYP_2C9_MolMechanics_optimized_QSARI.

Appendix S3. CYP_2C9_Quantum_optimized_QSARI.

Appendix S4. Dataset Information.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.