Predictive QSAR modeling of aldose reductase inhibitors using Monte Carlo feature selection

ARTICLE *in* EUROPEAN JOURNAL OF MEDICINAL CHEMISTRY · FEBRUARY 2014

Impact Factor: 3.45 · DOI: 10.1016/j.ejmech.2014.02.043 · Source: PubMed

6 AUTHORS, INCLUDING:

Chanin Nantasenamat
Mahidol University
82 PUBLICATIONS **985** CITATIONS
SEE PROFILE

Apilak Worachartcheewan
Mahidol University
56 PUBLICATIONS **471** CITATIONS
SEE PROFILE

Prasit Mandi
Mahidol University
13 PUBLICATIONS **74** CITATIONS
SEE PROFILE

Virapong Prachayasittikul
Mahidol University
200 PUBLICATIONS **1,863** CITATIONS
SEE PROFILE

European Journal of Medicinal Chemistry 76 (2014) 352–359

Contents lists available at ScienceDirect

European Journal of Medicinal Chemistry

journal homepage: http://www.elsevier.com/locate/ejmech

Original article

Predictive QSAR modeling of aldose reductase inhibitors using Monte Carlo feature selection

Chanin Nantasenamat [a,b,*], Teerawat Monnor [a], Apilak Worachartcheewan [a], Prasit Mandi [a,b], Chartchalerm Isarankura-Na-Ayudhya [b], Virapong Prachayasittikul [b]

[a] Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand
[b] Department of Clinical Microbiology and Applied Technology, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand

ARTICLE INFO

Article history:
Received 15 August 2013
Received in revised form
12 February 2014
Accepted 15 February 2014
Available online 16 February 2014

Keywords:
Aldose reductase
Aldose reductase inhibitor
QSAR
Monte Carlo
MC-MLR

ABSTRACT

This study explores the chemical space and quantitative structure–activity relationship (QSAR) of a set of 60 sulfonylpyridazinones with aldose reductase inhibitory activity. The physicochemical properties of the investigated compounds were described by a total of 3230 descriptors comprising of 6 quantum chemical descriptors and 3224 molecular descriptors. A subset of 5 descriptors was selected from the aforementioned pool by means of Monte Carlo (MC) feature selection coupled to multiple linear regression (MLR). Predictive QSAR models were then constructed by MLR, support vector machine and artificial neural network, which afforded good predictive performance as deduced from internal and external validation. The investigated models are capable of accounting for the origins of aldose reductase inhibitory activity and could be utilized in predicting this property in screening for novel and robust compounds.

1. Introduction

Diabetes mellitus is a pandemic disease affecting approximately 200 million people worldwide. It is a formidable condition that is expected to rise to 350 million by the year 2025 [1]. The polyol pathway has been implicated in the development of diabetes mellitus owing to its involvement in glucose metabolism. As such, a lucrative therapeutic approach is to modulate excess glucose flux by inhibiting two crucial steps of the polyol pathway namely through the rate-limiting conversion of glucose to sorbitol by aldose reductase (AR) and the subsequent conversion of sorbitol to fructose by sorbitol dehydrogenase. AR has been suggested as the etiological cause of long-term diabetic complications [2]. Particularly, it contributes to high concentrations of sorbitol that has been linked to the development of cataract, neuropathy, nephropathy and retinopathy. The inhibition of AR has thus been proposed as a viable approach in controlling diabetic complications.

Much effort has thus been invested in the search for novel aldose reductase inhibitors (ARI) as evident from a search in Scopus in which there were 1148 articles having "aldose reductase inhibitor*" as keywords in the article title prior to submission of this article (July 4, 2013) while during the revision of this article (February 12, 2014) that number increased to 1166. Of all those articles, there were 105 articles describing the synthesis of structurally diverse classes and scaffolds with AR inhibitory activities. Recent examples of novel scaffolds among others include luteolin [3], keto-pyrrolyl-difluorophenol [4], piplartine [5], quinoxalinone [6] and pyridylthiadiazine [7].

Quantitative structure–activity/property relationship (QSAR/QSPR) is a computational methodology for discerning the linkage between chemical structures of investigated compounds with their respective biological activity/chemical property. This is achieved by learning the inherent patterns hidden within the data set of interest by employing traditional and machine learning techniques. Predictive QSAR/QSPR modeling had been successfully employed to investigate a myriad of biological activities [8–19] and chemical properties [20–24]. A more in-depth coverage on the concepts of QSAR/QSPR can be found in previous review articles [25,26].

One of the early QSAR study on ARI was reported by Sun et al. [27] on a series of 45 indolebutanoic acids in which they performed 3D-QSAR studies by means of Comparative Molecular Similarity Indices Analysis where hydrophobicity and hydrogen bond donor/acceptor were used as descriptors to afford R^2 and Q^2 values of 0.934 and 0.557, respectively. Prabhakar et al. [28] subsequently performed a high-dimensional QSAR study of 48 flavones in which

* Corresponding author. Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand.
E-mail address: chanin.nan@mahidol.ac.th (C. Nantasenamat).

http://dx.doi.org/10.1016/j.ejmech.2014.02.043
0223-5234/

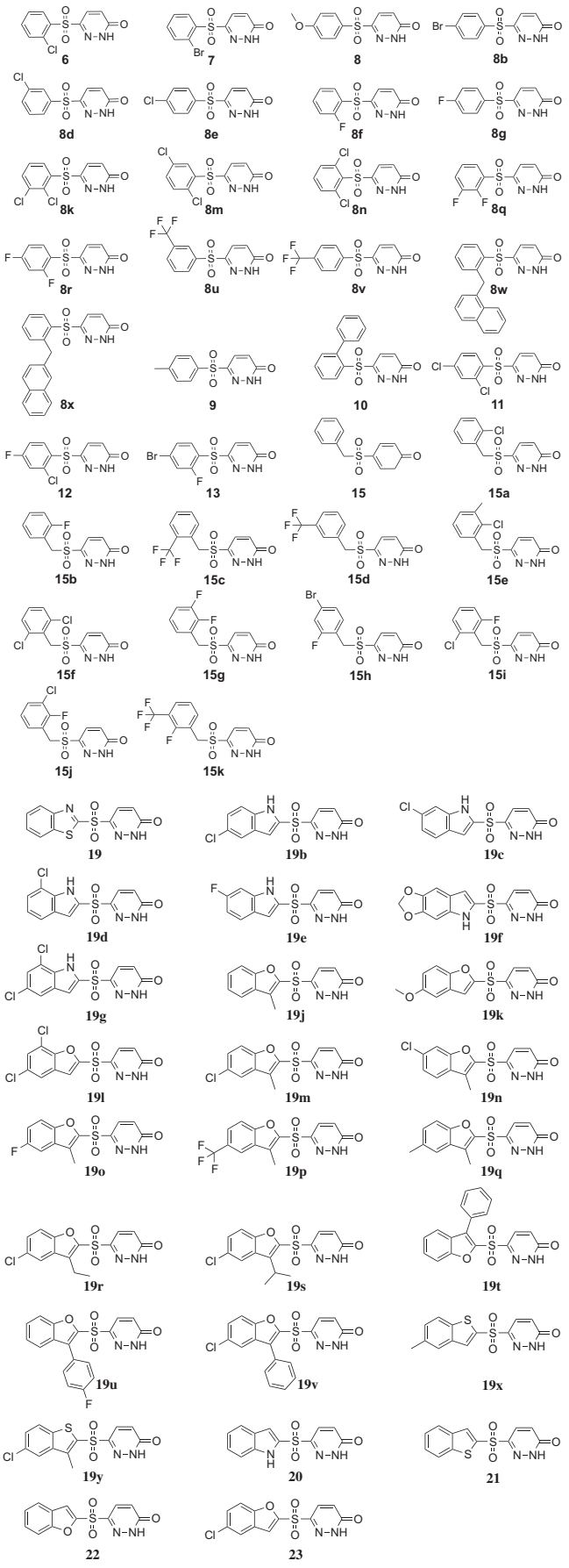

Fig. 1. Chemical structures of sulfonylpyridazinones.

they employed topological parameters (i.e., 152 Molconn-Z parameters and 6 indicators of flavone hydroxyls) as molecular descriptors while Free–Wilson, Combinatorial Protocol in Multiple Linear Regression (CP-MLR) and Partial Least Squares (PLS) were utilized for multivariate analysis. The best QSAR model was afforded by CP-MLR providing R^2 of 0.778 and Q^2 of 0.691 while PLS afforded slightly lower performance with R^2 of 0.740 and Q^2 of 0.671. Furthermore, Sambasivarao et al. [29] made use of a wide range of descriptors (i.e., thermodynamic, steric and electronic descriptors) to describe compounds in the set of 18 5-arylidene-2,4-thiazolidinediones while MLR was used as the multivariate analysis method to yield R^2 and Q^2 of 0.825 and 0.721, respectively. Hu et al. [30] employed two descriptors (i.e., electronegativity and molar volume) in predicting the logIC_{50} values of a set of 30 spirohydantoin derivatives using artificial neural network to produce R^2 of 0.822. Subsequently, Patra and co-workers [31,32] re-analyzed the previously mentioned data set of Hu et al. [30] in two separate studies using radial basis function network to give rise to Q^2_{CV} and Q^2_{Ext} of 0.647 and 0.891, respectively, as well as using artificial neural network to produce Q^2 value of 0.774. Thareja et al. [33] performed 3D-QSAR via self-organizing molecular field analysis on a series of 5-arylidine-2, 4-thiazolidinediones. Soni et al. [34] explored the use of QSAR modeling by means of MLR in correlating constitutional descriptor and 3D-MoRSE descriptors with AR inhibitory activity to afford R^2 and Q^2 of 0.880 and 0.788, respectively. Jain et al. [35] developed a QSAR model using physicochemical descriptors calculated in VLife MDS to yield model with R^2 and Q^2 of 0.75 and 0.62, respectively.

It can be seen that many of these QSAR models toward ARI were developed using a few specific classes of physicochemical descriptors in the absence of feature selection. Therefore, this study explores the use of Monte Carlo feature selection coupled to Multiple Linear Regression (MC-MLR) in choosing a subset of important variables from an initial pool of 3230 quantum chemical and molecular descriptors, which were derived from low energy conformers geometrically optimized at B3LYP/6-31G(d) level. After several rigorous calculations, the 5 most important descriptors were identified from the MC-MLR feature selection procedure. Such subset of descriptors was used in the construction of predictive QSAR models, which was demonstrated to afford good predictive performance.

2. Results and discussion

2.1. Exploring the chemical space of sulfonylpyridazinones

The investigated set of 60 sulfonylpyridazinones (Fig. 1) are AR inhibitors based on the non-hydantoin and non-carboxylic acid chemotypes. Particularly, aldose reductase inhibitors based on the hydantoin chemotype (i.e., sorbinil) are very weakly acidic ($pK_a > 8$) accounting for it being un-ionized at blood pH, which contributes to its efficient tissue penetration as well as being potent *in vivo* with broad spectrum of tissue activity. However, these chemotype have been reported to cause skin rash, hypersensitivity and liver toxicity. As for aldose reductase inhibitors based on the carboxylic acid chemotype (i.e., zopolrestat), they are significantly more acidic (pK_a ~3–4) than hydantoins rendering its ionized state at blood pH and its less potency *in vivo* with narrow spectrum of tissue activity than those of hydantoins [36,37]. It was found that all compounds passed the Lipinski's rule of 5 in which the molecular weight is <500 Da, LogP is <5, hydrogen bond donor is <5 and hydrogen bond acceptor is <10, thereby suggesting its drug-like properties. Navigation of the chemical space of sulfonylpyridazinones was performed in efforts to gain a clearer understanding of its structure–activity

relationship by analyzing the aforementioned set of descriptors from Lipinski's rule. Chemical space analysis provides essential knowledge on the general characteristics governing the biological activity of compounds [38].

Prior to chemical space analysis, the data set was stratified into 4 subsets as a function of their IC_{50} values in which subsets 1, 2, 3 and 4 span values of 1–92 nM, 118–280 nM, 350–667 nM as well as 870–3800 nM, respectively. Statistical analysis was then performed on the 4 subsets and it was observed that 3 of 4 Lipinski's descriptors (i.e., MW, ALogP and nHDon) as well as the IC_{50} were statistically significant with *P*-values less than 0.05 (Table 2). Visual representation of the overall distribution of data values of Lipinski's descriptors is shown as radar plots in Fig. 2. It can be seen that the MW and ALogP are inversely correlated with IC_{50} in which MW and ALogP increases as the IC_{50} decreases. On the other hand, nHDon is positively correlated with IC_{50} in which nHDon increases with increasing IC_{50}. As for the nHAcc descriptor, no significant differences were observed for the four subsets of compounds that were stratified by their IC_{50} values albeit the significantly lower nHAcc value in subset 1 (i.e., $\leq$100 nm) as compared to subset 2 through 4 (i.e., 101–300 nm, 301–700 nm and $\geq$700 nm, respectively).

Among the investigated compounds, **19m** was the most potent affording IC_{50} value of 1 nM that was significantly higher than the parent compound **19** having IC_{50} of 450 nM. Furthermore, a closer look at the chemical structures indicated that the four most potent compounds **19m** along with **19o**, **19p** and **19r** had oxygen at the X position, halogens (i.e., Cl, F and CF_3) at the R_3 position and CH_3 at the R_5 position. The four least potent compounds (i.e., **19f**, **10**, **8** and **9**) had IC_{50} values in the range of 1900–3800 nM and a closer look at the chemical structures revealed the presence of dioxolo (i.e., OCH_2O) bridging R_2 and R_3, phenyl at R_5, OCH_3 at R_3 and CH_3 at R_3 positions, respectively. It is interesting to note that substituents for the other compounds of this structural class (i.e., **6**, **7**, **11**–**13** and the **8** series of compounds) had halogens (i.e., Br, Cl and F) at the mentioned positions whereas the former did not. Taken together, it can be seen that halogens are essential for the observed AR inhibitory activity.

Table 1
Description of molecular descriptors.

Descriptor	Class	Definition
F04[C–N]	2D frequency fingerprints	Frequency of C–N at topological distance 04
ATS4m	2D autocorrelations	Broto-Moreau autocorrelation of a topological structure – lag 4/ weighted by atomic masses
B05[O–Cl]	2D binary fingerprints	Presence/absence of O–Cl at topological distance 05
Mor14e	3D-MoRSE descriptors	3D-MoRSE – signal 14/weighted by atomic Sanderson electronegativities
qpmax	Charge descriptors	Maximum positive charge

2.2. Feature selection of descriptors

MC-MLR was employed to select important descriptors from the initial pool of 3230 descriptors. Important descriptors were identified as those that frequently occur in QSAR models passing the Q^2 threshold. The 5 frequently occurring features included F04[C–N], ATS4m, B05[O–Cl], Mor14e and qpmax, which were further employed in the construction of QSAR models. Definitions of significant descriptors are shown in Table 1. It can be seen that 2 of the 5 descriptors are topological distances between C–N and O–Cl atom pairs (i.e., F04[C–N] and B05[O–Cl]). The next 2 of 5 descriptors are uniform-length descriptors derived by encoding a set of molecular properties via autocorrelation, which in this case pertains to atomic masses and electronegativities (i.e., ATS4m and Mor14e). The last descriptor describes the maximum positive charge of a molecule. It is interesting to note that one of the descriptor particularly B05[O–Cl] pertained to the halogen Cl atom, which as previously noted was crucial for AR inhibitory activity.

It should also be noted that 2 descriptors are topological descriptors while 3 descriptors are geometrical descriptors derived directly from the xyz coordinates. The former essentially constitute descriptors that are simple and inexpensive to calculate while the latter provide rather large information content albeit at the cost of a heavier computation that is required in optimizing the molecular structure.

2.3. Development of QSAR model using MLR

A subset comprising of 54 compounds or 90% from the full set of 60 compounds was used as the training set and 10-fold cross-validation set for QSAR model development using MLR. A second subset containing 6 compounds (i.e., **6**, **8**, **8b**, **12**, **19g** and **19t**) or 10% from the full set of 60 compounds were used as the external testing set. The chemical space spanned by both sets of data is shown in Fig. 3 in which it can be seen that the external testing set was in the domain of applicability covered by the internal set.

The set of 5 important descriptors as selected by MC-MLR were used as independent variables while aldose reductase inhibitory activity (i.e., pIC_{50}) was used as the dependent variable. Outlying compounds were identified as those having standardized residuals greater than 2 SD units. The MLR method generated 6 models and 12 compounds (i.e., **10**, **20**, **15c**, **15h**, **19k**, **19l**, **19m**, **19n**, **19o**, **19p**, **19q** and **19r**) were detected as outliers and removed from the data set. The predictive performance of the constructed MLR models is shown in Table 3. It can be seen that MLR model 6 afforded the best performance. The MLR equation of this model along with its statistical parameters is shown below.

Table 2
Statistical comparison of Lipinski's descriptors for compounds falling into 4 activity ranges.

	Activity ranges[a]				*P*-value
	1	2	3	4	
MW	328.93 ± 24.86	304.47 ± 27.96	294.64 ± 36.21	289.08 ± 26.48	0.002[b]
ALogP	2.57 ± 0.50	2.06 ± 0.62	1.78 ± 0.69	1.63 ± 0.53	<0.001[c]
nHDon	1.00 ± 0.00	1.06 ± 0.24	1.20 ± 0.41	1.43 ± 0.53	0.013[c]
nHAcc	6.19 ± 1.21	5.71 ± 0.92	5.87 ± 1.06	5.86 ± 0.90	0.563[c]
IC_{50}	40.46 ± 27.18	187.06 ± 44.76	478.53 ± 107.46	1895.71 ± 971.31	<0.001[c]

Data expressed as Mean ± SD.
[a] 1, 2, 3 and 4 corresponds to IC_{50} values of $\leq$100, 101–300, 301–700 and $\geq$700 nm.
[b] Statistical significance test was performed using ANOVA test.
[c] Statistical significance test was performed using Kruskal–Wallis H test.

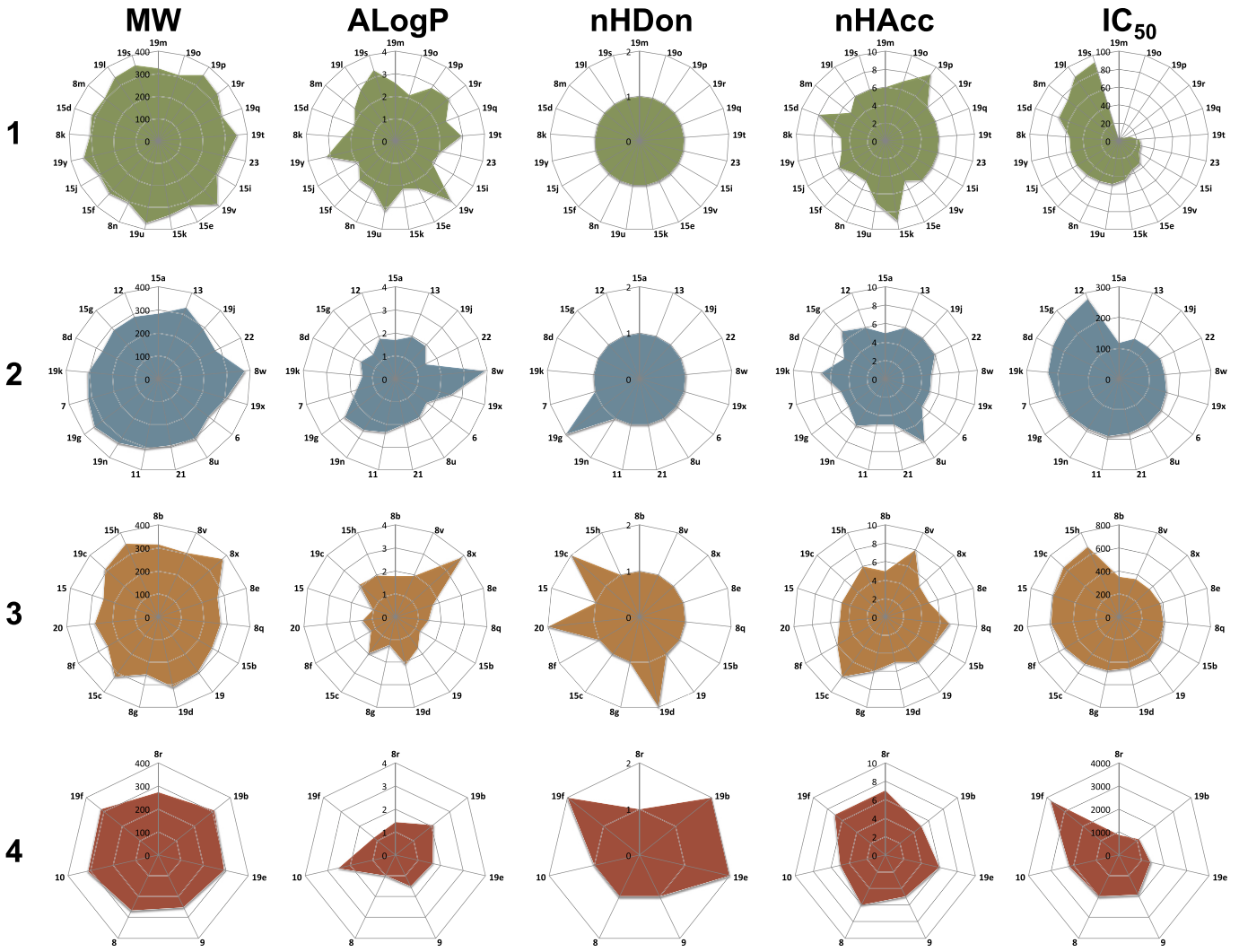

Fig. 2. Radar plots of Lipinski's descriptors and IC_{50} values for the four subsets of data (1–4) representing compounds having IC_{50} values in the range of 1–92 nM, 118–280 nM, 350–667 nM as well as 870–3800 nM, respectively.

$$\begin{aligned} pIC_{50} &= -0.332(F04[C{-}N]) + 0.188(ATS4m) \\ &\quad + 0.097(B05[O{-}Cl]) - 0.114(Mor14e) \\ &\quad - 0.107(qpmax) + 6.766 \end{aligned} \tag{1}$$

$N = 42$, $R^2_{Tr} = 0.849$, $RMSE_{Tr} = 0.204$, $Q^2_{CV} = 0.794$, $RMSE_{CV} = 0.244$, $F{-}ratio_{CV} = 27.753$ ($F_{Critical} = 2.477$), $Q^2_{Ext} = 0.741$, $RMSE_{Ext} = 0.325$

It can be seen from the regression coefficients that the most important molecular descriptors were F04[C–N] > ATS4m > Mor14e > qpmax > B05[O–Cl] which displayed corresponding values of −0.332, 0.188, −0.114, −0.107, 0.097, respectively.

Plot of experimental versus predicted activities for investigated compounds is shown in Fig. 4a.

2.4. Development of QSAR model using ANN

In addition to MLR and SVM, the 54 compound data set was also used in the development of QSAR models using ANN. Default parameters of ANN as set by the Weka software is comprised of 5 hidden nodes, learning rate of 0.3, momentum of 0.2 and learning epoch size of 500. Such parameters were used for outlier detection and removal. ANN models 1 through 9 were obtained and 18 compounds (i.e., **7**, **9**, **15**, **19**, **15c**, **15f**, **15h**, **19k**, **19l**, **19m**, **19n**, **19o**, **19p**, **19q**, **19r**, **19v**, **8u** and **8w**) were detected as outliers and removed from the data set. ANN model 9 were then subjected to parameter optimization to yield the following: hidden node, learning epoch, learning rate and momentum of 3, 300, 0.1 and 0.7, respectively, which gave rise to ANN model 10. As shown in Table 4, the results indicated that the final ANN model provided good predictive performance with the following parameters: $R_{Tr} = 0.949$ and $RMSE_{Tr} = 0.152$ for the training set; $Q^2_{CV} = 0.854$, $RMSE_{CV} = 0.216$ and F-ratio = 35.096 for the 10-fold CV set; and $Q^2_{Ext} = 0.624$ and $RMSE_{Ext} = 0.384$ for the external test set. Plot of experimental versus predicted activities of investigated compounds as predicted by ANN is shown in Fig. 4b.

2.5. Development of QSAR model using SVM

The set of 54 compounds were used in the construction of QSAR models using SVM. The input space of the SVM model is

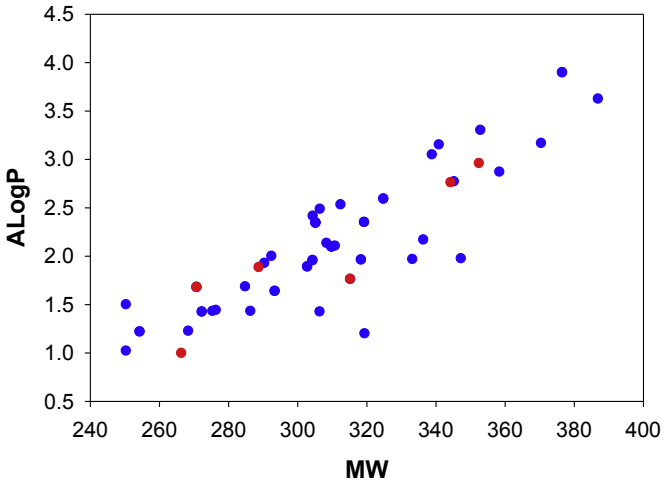

Fig. 3. Chemical space of internal and external subsets of data are shown as blue and red dots, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

transformed via the radial basis function kernel to a higher dimensional feature space in which maximal hyperplanes for distinguishing the modeled property (i.e., AR inhibitory activity). SVM model building was initiated by searching for optimal *C* and γ parameters via two-level grid search comprising of the initial global grid search followed by a refined local grid search. Outliers were identified and discarded from the model. It was found that the optimal C and γ values from the global grid search were 2^{13} and 2^{-9}, respectively, while the refined local grid search afforded optimal C and γ values of $2^{12.7}$ and $2^{-8.9}$, respectively. The predictive performance of SVM models 1 through 9 is shown in Table 4. The results indicated that 15 compounds (i.e., **10, 20, 15c, 15h, 15j, 19f, 19l, 19m, 19n, 19o, 19p, 19q, 19r, 8e** and **8g)** were identified as outliers and subjected to removal from the data set. SVM parameter optimization was performed once again on the new data set and the optimal C and γ parameters as elucidated from global grid search had values of 2^{9} and 2^{-7}, respectively, while optimal C and γ values as obtained from local grid search were $2^{9.8}$ and $2^{-7.2}$, respectively (SVM model 10). Statistical parameters of SVM models 1–10 are presented in Table 5. The best performing SVM model 10 provided the following statistical parameters: $R^2_{Tr} = 0.894$ and $RMSE_{Tr} = 0.160$ for the training set; $Q^2_{CV} = 0.864$, $RMSE_{CV} = 0.180$ and *F*-ratio = 41.929 for the 10-fold CV set; and $Q^2_{Ext} = 0.788$ and $RMSE_{Ext} = 0.293$ for the external test set. Plot of experimental versus predicted activities of the investigated compounds as predicted by SVM is shown in Fig. 4c.

As can be seen, the results indicated that all multivariate methods provided good performance in predicting the pIC_{50} values of the investigated sulfonylpyridazinones. This is verified by both internal and external validations of the predictive QSAR models.

3. Conclusion

In summary, this study aims to establish formal relationship between physicochemical features of investigated compounds based on sulfonylpyridazinones and their respective AR inhibitory activity. Feature selection via the MC-MLR method identified 5 important descriptors that were selected from a pool of 3230 descriptors. The selected descriptors indicated the importance of topological distances at C–N and O–Cl atom pairs, uniform-length descriptors pertaining to atomic masses and electronegativities as well as the maximum positive charge of molecules in accounting for the origins of AR inhibitory activities. The QSAR methodologies presented herein could further be used for the design of novel AR inhibitors for the treatment of diabetes.

4. Material and methods

4.1. Data set

A data set comprising of 60 compounds based on sulfonylpyridazinones was obtained from the work of Mylari et al. [36]. Such non-carboxylic acid, non-hydantoin inhibitors of aldose reductase was tested *in vitro* using human recombinant aldose reductase. Inhibition was determined from the percentage reduction in NADPH oxidation rate as compared to samples in absence of the investigated compounds. The bioactivity of aldose reductase is expressed by their inhibitory concentration at 50% (IC_{50}).

4.2. Geometry optimization

The initial chemical structures of the investigated compounds were drawn using MarvinSketch and converted to three-dimensional structure with explicit hydrogens added using the command-line *molconvert* tool of MarvinSketch. The molecular structures were then saved in the appropriate input format for further geometry optimization using Gaussian 09 at the density functional theory level using Becke's three-parameter Lee-Yang-Parr (B3LYP) with the 6-31G(d) basis set.

4.3. Molecular descriptors calculation

A set of 6 quantum chemical descriptors was obtained from low energy conformers as calculated by Gaussian 09 [39] at B3LYP/6-31G(d) level. This set is comprised of the following descriptors: total energy of a molecule (Energy), mean absolute charge (Q_m), dipole moment, energy of the highest occupied molecular orbital (HOMO), energy of the lowest unoccupied molecular orbital (LUMO) and the energetic difference of HOMO and LUMO (HOMO-LUMO). A set of 3224 molecular descriptors were calculated using Dragon [40] that spans the following 22 descriptor classes: constitutional descriptors, topological descriptors, walk and path counts, connectivity indices, information indices, 2D autocorrelations, edge adjacency indices, Burden eigenvalues, topological charge indices, eigenvalue-based

Table 3
Summary of predictive performance of QSAR models developed using MLR method.

Model	*N*	R^2_{Tr}	$RMSE_{Tr}$	Q^2_{CV}	$RMSE_{CV}$	Q^2_{Ext}	$RMSE_{Ext}$	*F*-ratio$_{CV}$	$F_{Critical}$
1	54	0.598	0.451	0.494	0.513	0.646	0.363	9.376	2.409
2	51	0.697	0.322	0.617	0.348	0.561	0.395	14.489	2.422
3	47	0.775	0.262	0.710	0.292	0.662	0.361	20.086	2.443
4	45	0.815	0.231	0.752	0.274	0.687	0.354	23.636	2.456
5	43	0.836	0.212	0.776	0.249	0.742	0.330	25.674	2.470
6	42	0.849	0.204	0.794	0.244	0.741	0.325	27.753	2.477

indices, Randic molecular profiles, geometrical descriptors, RDF descriptors, 3D-MoRSE descriptors, WHIM descriptors, GETAWAY descriptors, functional group counts, atom-centered fragments, charge descriptors, molecular properties, 2D binary fingerprints and 2D frequency fingerprints.

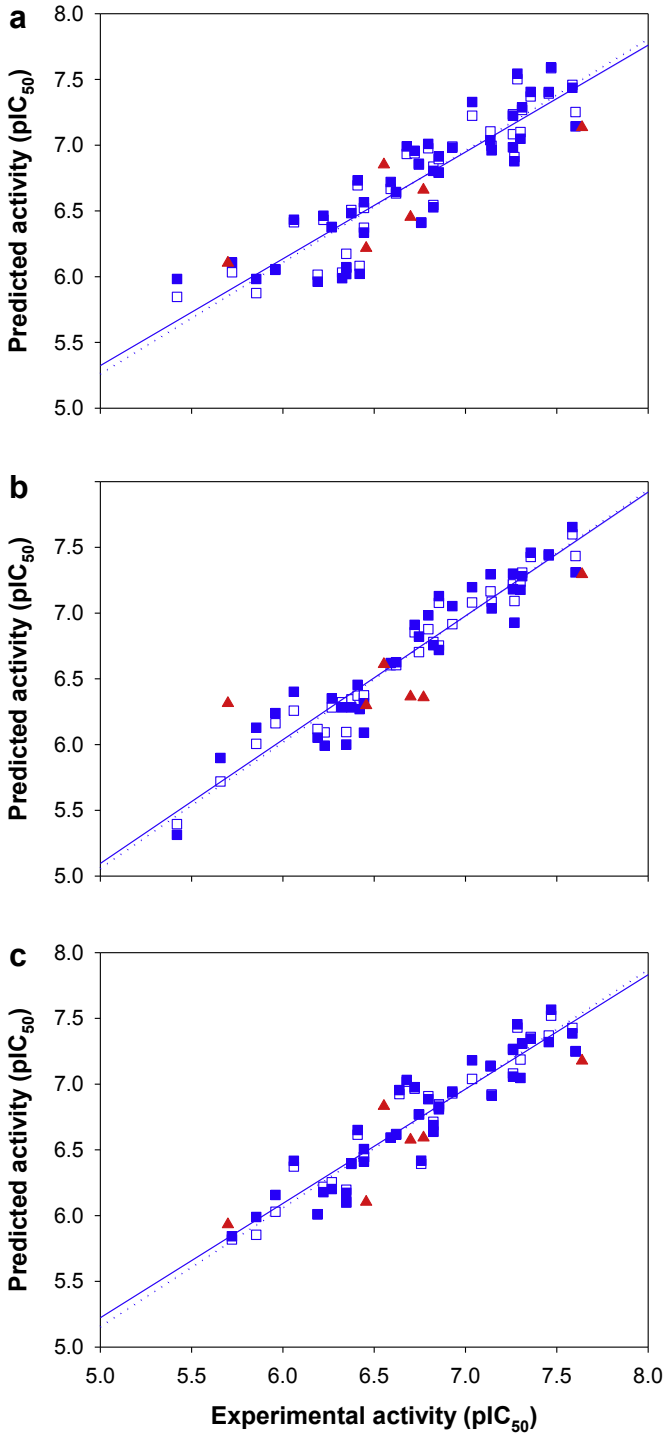

Fig. 4. Plot of experimental versus predicted aldose inhibitory activity using MLR (a), ANN (b) and SVM (c). Data samples and trend line are shown as blue closed squares and solid line respectively, for the leave-one-out cross-validation set, blue open squares and dotted line, respectively, are used in representing the training set while red closed triangles represents the external test set. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.4. Model development

4.4.1. Data pre-processing and data sampling

The following data pre-processing procedures were performed: (i) constant variables with standard deviation less than 0.10 were removed to yield 1292 descriptors, (ii) descriptors were then standardized, (iii) pairwise correlation coefficient values were computed for all possible descriptor pairs and those having inter-correlation values greater than 0.70 were deemed highly correlated and redundant warranting its removal to yield a set of 112 descriptors. The data set was then randomly divided into two portion comprising of 90% and 10% of the data set for use as the internal cross-validation set and external test set, respectively. Ten-fold cross-validation was performed in which one set was left out as the testing set while the remaining sets were used in training a predictive model. This was iteratively performed so that all sets had a chance to be used as the testing set. The external test set serves to verify the predictive power of the QSAR models.

4.4.2. Monte Carlo-based feature selection

Feature selection is a crucial step in multivariate analysis as it allows the identification of important subset of molecular descriptors. In this respect, our study employs Monte Carlo as a searching method in which a subset of descriptors was selected from the initial pool of descriptors. Fitness of the selected subset of descriptors was determined from its predictive performance that was made possible by coupling it to multiple linear regression. The Monte Carlo-based feature selection algorithm was coded in Python.

Briefly, the MC-MLR algorithm entails the selection of an initial subset of N descriptors (where N equals 3, 4, 5, 6, 7, etc.) in a stochastic manner from the pool of 112 descriptors. The predictive performance of the descriptor subset was assessed from Q^2 values obtained from 10-fold cross-validation of MLR modeling. Subsets affording values greater than the desired Q^2 threshold of 0.5 is accepted whereas subsets affording a lesser value are rejected and a variant subset are created by randomly replacing one descriptor in the subset. This cycle was performed for 20,000 iterations as to generate a relatively large sum of subset variants. The relative importance of descriptors was deduced as highly prevalent descriptors obtained from these satisfactory models. This led to the identification of the 5 most important descriptors that were used in subsequent multivariate analysis.

4.4.3. Multivariate analysis

A subset of 5 important descriptors was then employed in modeling AR inhibitory activities using various multivariate analysis methods including MLR, support vector machine (SVM) and artificial neural network (ANN). MLR was calculated under the Python environment. ANN and SVM were calculated using the back-propagation and sequential minimal optimization algorithms, respectively, from the Weka data mining package [41]. Parameter optimizations of ANN and SVM parameters were performed in order to obtain the best performing parameters for QSAR model calculations. The following ANN parameters were optimized: the number of hidden nodes, learning epochs, learning rate and momentum. The following SVM parameters were subjected to optimizations: C and γ.

4.5. Statistical analysis

Significance test was performed using SPSS version 18.0 on Lipinski's descriptors for the 4 subsets of compounds that were stratified by their IC_{50} activity. Normality of each data subsets was assessed using Kolmogorov–Smirnov test. Difference of non-

Table 4
Summary of predictive performance of QSAR models developed using ANN method.

Model	N	R^2_{Tr}	$RMSE_{Tr}$	Q^2_{CV}	$RMSE_{CV}$	Q^2_{Ext}	$RMSE_{Ext}$	F-ratio$_{CV}$	$F_{Critical}$
1	54	0.671	0.398	0.259	0.773	0.395	0.655	3.355	2.409
2	51	0.785	0.396	0.391	0.615	0.688	0.406	5.778	2.422
3	47	0.825	0.278	0.524	0.470	0.630	0.392	9.027	2.443
4	45	0.880	0.242	0.635	0.391	0.569	0.437	13.570	2.456
5	44	0.875	0.227	0.645	0.375	0.682	0.368	13.808	2.463
6	40	0.909	0.185	0.768	0.270	0.647	0.383	22.510	2.494
7	39	0.936	0.142	0.760	0.273	0.659	0.372	20.900	2.503
8	37	0.946	0.152	0.825	0.239	0.596	0.404	29.229	2.523
9	36	0.945	0.163	0.827	0.236	0.591	0.402	28.682	2.534
10	36	0.949	0.152	0.854	0.216	0.624	0.384	35.096	2.534

Table 5
Summary of predictive performance of QSAR models developed using SVM method.

Model	N	R^2_{Tr}	$RMSE_{Tr}$	Q^2_{CV}	$RMSE_{CV}$	Q^2_{Ext}	$RMSE_{Ext}$	F-ratio$_{CV}$	$F_{Critical}$
1	54	0.586	0.465	0.543	0.484	0.773	0.300	11.407	2.409
2	51	0.690	0.327	0.603	0.372	0.746	0.315	13.670	2.422
3	48	0.760	0.263	0.674	0.313	0.748	0.314	17.367	2.438
4	46	0.768	0.252	0.708	0.284	0.792	0.299	19.397	2.449
5	44	0.814	0.209	0.758	0.242	0.783	0.294	23.805	2.463
6	43	0.826	0.206	0.791	0.224	0.793	0.295	28.007	2.470
7	41	0.862	0.185	0.822	0.208	0.790	0.295	32.326	2.485
8	40	0.877	0.173	0.854	0.187	0.786	0.294	39.775	2.494
9	39	0.891	0.161	0.860	0.183	0.785	0.295	40.543	2.503
10	39	0.894	0.160	0.864	0.180	0.788	0.293	41.929	2.503

normal distribution was assessed using Kruskal–Wallis H test while difference of normal distribution was determined using ANOVA test.

The predictive performance of the QSAR models were evaluated from the squared correlation coefficient value from the 10-fold cross-validated set (Q^2) in comparison with correlation coefficient value obtained from the training set (R^2). Furthermore, the prediction error can be deduced from its root mean squared error (RMSE) value. Moreover, the F ratio between the explained and unexplained variance having m and $(n - m - 1)$ degrees of freedom was calculated as follows:

$$F_{(m,n-m-1)} = \frac{Q^2/m}{(1-Q^2)/(n-m-1)} \quad (2)$$

where m represents the number of molecular descriptors and n denotes the number of compounds in the data set.

Taken together, highly predictive models are those having high Q^2, low RMSE and high F ratio.

Acknowledgments

This project is supported by the annual budget grant (B.E. 2556-2558) of Mahidol University.

References

[1] S. Wild, G. Roglic, A. Green, R. Sicree, H. King, Global prevalence of diabetes: estimates for the year 2000 and projections for 2030, Diabetes Care 27 (2004) 1047–1053.
[2] J.H. Kinoshita, C. Nishimura, The involvement of aldose reductase in diabetic complications, Diabetes Metabolism Reviews 4 (1988) 323–337.
[3] Q.Q. Wang, N. Cheng, X.W. Zheng, S.M. Peng, X.Q. Zou, Synthesis of organic nitrates of luteolin as a novel class of potent aldose reductase inhibitors, Bioorganic & Medicinal Chemistry 21 (2013) 4301–4310.
[4] E. Kotsampasakou, V.J. Demopoulos, Synthesis of derivatives of the keto-pyrrolyl-difluorophenol scaffold: some structural aspects for aldose reductase inhibitory activity and selectivity, Bioorganic & Medicinal Chemistry 21 (2013) 869–873.
[5] V.R. Rao, P. Muthenna, G. Shankaraiah, C. Akileshwari, K.H. Babu, G. Suresh, K.S. Babu, R.S. Chandra Kumar, K.R. Prasad, P.A. Yadav, J.M. Petrash, G.B. Reddy, J.M. Rao, Synthesis and biological evaluation of new piplartine analogues as potent aldose reductase inhibitors (ARIs), European Journal of Medicinal Chemistry 57 (2012) 344–361.
[6] Y. Yang, S. Zhang, B. Wu, M. Ma, X. Chen, X. Qin, M. He, S. Hussain, C. Jing, B. Ma, C. Zhu, An efficient synthesis of quinoxalinone derivatives as potent inhibitors of aldose reductase, ChemMedChem 7 (2012) 823–835.
[7] X. Chen, Y. Yang, B. Ma, S. Zhang, M. He, D. Gui, S. Hussain, C. Jing, C. Zhu, Q. Yu, Y. Liu, Design and synthesis of potent and selective aldose reductase inhibitors based on pyridylthiadiazine scaffold, European Journal of Medicinal Chemistry 46 (2011) 1536–1544.
[8] P. Mandi, C. Nantasenamat, K. Srungboonmee, C. Isarankura-Na-Ayudhya, V. Prachayasittikul, QSAR study of anti-prion activity of 2-aminothiazoles, EXCLI Journal 11 (2012) 453–467.
[9] C. Nantasenamat, C. Isarankura-Na-Ayudhya, T. Naenna, V. Prachayasittikul, Prediction of bond dissociation enthalpy of antioxidant phenols by support vector machine, Journal of Molecular Graphics and Modelling 27 (2008) 188–196.
[10] C. Nantasenamat, T. Piacham, T. Tantimongcolwat, T. Naenna, C. Isarankura-Na-Ayudhya, V. Prachayasittikul, QSAR model of the quorum-quenching *N*-acyl-homoserine lactone lactonase activity, Journal of Biological Systems 16 (2008) 279–293.
[11] R. Pingaew, P. Tongraung, A. Worachartcheewan, C. Nantasenamat, S. Prachayasittikul, S. Ruchirawat, V. Prachayasittikul, Cytotoxicity and QSAR study of (thio)ureas derived from phenylalkylamines and pyridylalkylamines, Medicinal Chemistry Research 22 (2013) 4016–4029.
[12] C. Thippakorn, T. Suksrichavalit, C. Nantasenamat, T. Tantimongcolwat, C. Isarankura-Na-Ayudhya, T. Naenna, V. Prachayasittikul, Modeling the LPS neutralization activity of anti-endotoxins, Molecules 14 (2009) 1869–1888.
[13] A. Worachartcheewan, C. Nantasenamat, C. Isarankura-Na-Ayudhya, S. Prachayasittikul, V. Prachayasittikul, Predicting the free radical scavenging activity of curcumin derivatives, Chemometrics and Intelligent Laboratory Systems 109 (2011) 207–216.
[14] A. Worachartcheewan, C. Nantasenamat, C. Isarankura-Na-Ayudhya, V. Prachayasittikul, Predicting antimicrobial activities of benzimidazole derivatives, Medicinal Chemistry Research 22 (2013) 5418–5430.
[15] A. Worachartcheewan, C. Nantasenamat, T. Naenna, C. Isarankura-Na-Ayudhya, V. Prachayasittikul, Modeling the activity of furin inhibitors using artificial neural network, European Journal of Medicinal Chemistry 44 (2009) 1664–1673.
[16] A. Worachartcheewan, C. Nantasenamat, W. Owasirikul, T. Monnor, O. Naruepantawart, S. Janyapaisarn, S. Prachayasittikul, V. Prachayasittikul, Insights into antioxidant activity of 1-adamantylthiopyridine analogs using multiple linear regression, European Journal of Medicinal Chemistry 73 (2014) 258–264.
[17] C. Nantasenamat, A. Worachartcheewan, P. Mandi, T. Monnor, C. Isarankura-Na-Ayudhya, V. Prachayasittikul, QSAR modeling of aromatase inhibition by flavonoids using machine learning approaches, Chemical Papers 68 (2014) 697–713.

[18] C. Nantasenamat, A. Worachartcheewan, S. Prachayasittikul, C. Isarankura-Na-Ayudhya, V. Prachayasittikul, QSAR modeling of aromatase inhibitory activity of 1-substituted 1,2,3-triazole analogs of letrozole, European Journal of Medicinal Chemistry 69 (2013) 99–114.
[19] A. Worachartcheewan, C. Nantasenamat, C. Isarankura-Na-Ayudhya, V. Prachayasittikul, QSAR study of amidino bis-benzimidazole derivatives as potent anti-malarial agents against *Plasmodium falciparum*, Chemical Papers 67 (2013) 1462–1473.
[20] C. Nantasenamat, C. Isarankura-Na-Ayudhya, T. Naenna, V. Prachayasittikul, Quantitative structure-imprinting factor relationship of molecularly imprinted polymers, Biosensors & Bioelectronics 22 (2007) 3309–3317.
[21] C. Nantasenamat, C. Isarankura-Na-Ayudhya, N. Tansila, T. Naenna, V. Prachayasittikul, Prediction of GFP spectral properties using artificial neural network, Journal of Computational Chemistry 28 (2007) 1275–1289.
[22] C. Nantasenamat, T. Naenna, C. Isarankura Na-Ayudhya, V. Prachayasittikul, Quantitative prediction of imprinting factor of molecularly imprinted polymers by artificial neural network, Journal of Computer-Aided Molecular Design 19 (2005) 509–524.
[23] C. Nantasenamat, K. Srungboonmee, S. Jamsak, N. Tansila, C. Isarankura-Na-Ayudhya, V. Prachayasittikul, Quantitative structure-property relationship study of spectral properties of green fluorescent protein with support vector machine, Chemometrics and Intelligent Laboratory Systems 120 (2013) 42–52.
[24] M. Lapins, A. Worachartcheewan, O. Spjuth, V. Georgiev, V. Prachayasittikul, C. Nantasenamat, J.E. Wikberg, A unified proteochemometric model for prediction of inhibition of cytochrome p450 isoforms, PLoS One 8 (2013) e66566.
[25] C. Nantasenamat, C. Isarankura-Na-Ayudhya, T. Naenna, V. Prachayasittikul, A practical overview of quantitative structure-activity relationship, EXCLI Journal 8 (2009) 74–88.
[26] C. Nantasenamat, C. Isarankura-Na-Ayudhya, V. Prachayasittikul, Advances in computational methods to predict the biological activity of compounds, Expert Opinion on Drug Discovery 5 (2010) 633–654.
[27] W.S. Sun, Y.S. Park, J. Yoo, K.D. Park, S.H. Kim, J.H. Kim, H.J. Park, Rational design of an indolebutanoic acid derivative as a novel aldose reductase inhibitor based on docking and 3D QSAR studies of phenethylamine derivatives, Journal of Medicinal Chemistry 46 (2003) 5619–5627.
[28] Y.S. Prabhakar, M.K. Gupta, N. Roy, Y. Venkateswarlu, A high dimensional QSAR study on the aldose reductase inhibitory activity of some flavones: topological descriptors in modeling the activity, Journal of Chemical Information and Modeling 46 (2006) 86–92.
[29] S.V. Sambasivarao, L.K. Soni, A.K. Gupta, P. Hanumantharao, S.G. Kaskhedikar, Quantitative structure-activity analysis of 5-arylidene-2,4-thiazolidinediones as aldose reductase inhibitors, Bioorganic & Medicinal Chemistry Letters 16 (2006) 512–520.
[30] L. Hu, G. Chen, R.M. Chau, A neural networks-based drug discovery approach and its application for designing aldose reductase inhibitors, Journal of Molecular Graphics and Modelling 24 (2006) 244–253.
[31] J.C. Patra, B.H. Chua, Artificial neural network-based drug design for diabetes mellitus using flavonoids, Journal of Computational Chemistry 32 (2011) 555–567.
[32] J.C. Patra, O. Singh, Artificial neural networks-based approach to design ARIs using QSAR for diabetes mellitus, Journal of Computational Chemistry 30 (2009) 2494–2508.
[33] S. Thareja, S. Aggarwal, T.R. Bhardwaj, M. Kumar, 3D-QSAR studies on a series of 5-arylidine-2, 4-thiazolidinediones as aldose reductase inhibitors: a self-organizing molecular field analysis approach, Journal of Medicinal Chemistry 6 (2010) 30–36.
[34] L.K. Soni, A.K. Gupta, S.G. Kaskhedikar, Exploration of QSAR modelling techniques and their combination to rationalize the physicochemical characters of nitrophenyl derivatives towards aldose reductase inhibition, Journal of Enzyme Inhibition and Medicinal Chemistry 24 (2009) 1002–1007.
[35] S.V. Jain, K.S. Bhadoriya, S.B. Bari, QSAR and flexible docking studies of some aldose reductase inhibitors obtained from natural origin, Medicinal Chemistry Research 21 (2012) 1665–1676.
[36] B.L. Mylari, S.J. Armento, D.A. Beebe, E.L. Conn, J.B. Coutcher, M.S. Dina, M.T. O'Gorman, M.C. Linhares, W.H. Martin, P.J. Oates, D.A. Tess, G.J. Withbroe, W.J. Zembrowski, A novel series of non-carboxylic acid, non-hydantoin inhibitors of aldose reductase with potent oral activity in diabetic rat models: 6-(5-chloro-3-methylbenzofuran-2-sulfonyl)-2H-pyridazin-3-one and congeners, Journal of Medicinal Chemistry 48 (2005) 6326–6339.
[37] B.L. Mylari, S.J. Armento, D.A. Beebe, E.L. Conn, J.B. Coutcher, M.S. Dina, M.T. O'Gorman, M.C. Linhares, W.H. Martin, P.J. Oates, D.A. Tess, G.J. Withbroe, W.J. Zembrowski, A highly selective, non-hydantoin, non-carboxylic acid inhibitor of aldose reductase with potent oral activity in diabetic rat models: 6-(5-chloro-3-methylbenzofuran- 2-sulfonyl)-2-H-pyridazin-3-one, Journal of Medicinal Chemistry 46 (2003) 2283–2286.
[38] C. Nantasenamat, H. Li, P. Mandi, A. Worachartcheewan, T. Monnor, C. Isarankura-Na-Ayudhya, V. Prachayasittikul, Exploring the chemical space of aromatase inhibitors, Molecular Diversity 17 (2013) 661–677.
[39] M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G.A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H.P. Hratchian, A.F. Izmaylov, J. Bloino, G. Zheng, J.L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J.A. Montgomery, J.E. Peralta, F. Ogliaro, M. Bearpark, J.J. Heyd, E. Brothers, K.N. Kudin, V.N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J.C. Burant, S.S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J.M. Millam, M. Klene, J.E. Knox, J.B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R.E. Stratmann, O. Yazyev, A.J. Austin, R. Cammi, C. Pomelli, J.W. Ochterski, R.L. Martin, K. Morokuma, V.G. Zakrzewski, G.A. Voth, P. Salvador, J.J. Dannenberg, S. Dapprich, A.D. Daniels, Farkas, J.B. Foresman, J.V. Ortiz, J. Cioslowski, D.J. Fox, Gaussian 09, Revision A.1, 2009. Wallingford, Connecticut.
[40] Talete srl, DRAGON for Windows (Software for Molecular Descriptor Calculations), Version 5.5, 2007. Milano, Italy.
[41] I.H. Witten, E. Frank, M.A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, third ed., Morgan Kaufmann, Amsterdam, Netherlands, 2011.