Chemometrics and Intelligent Laboratory Systems 138 (2014) 120–126

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Large-scale QSAR study of aromatase inhibitors using SMILES-based descriptors

Apilak Worachartcheewan [a,b], Prasit Mandi [a], Virapong Prachayasittikul [c], Alla P. Toropova [d], Andrey A. Toropov [d], Chanin Nantasenamat [a,c,*]

[a] Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand
[b] Department of Clinical Chemistry, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand
[c] Department of Clinical Microbiology and Applied Technology, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand
[d] IRCCS-Istituto di Ricerche Farmacologiche Mario Negri, Via La Masa 19, Milano, Italy

ARTICLE INFO

Article history:
Received 20 May 2014
Received in revised form 22 July 2014
Accepted 22 July 2014
Available online 7 August 2014

Keywords:
Aromatase inhibitors
Breast cancer
QSAR
SMILES
CORAL software
Data mining

ABSTRACT

Aromatase inhibitors (AIs) represent a promising therapeutic class of anticancer agents against estrogen receptor-positive breast cancer. Bioactivity data on pIC_{50} of 973 AIs were employed in the construction of quantitative structure-activity relationship (QSAR) models using COR relation And Logic (CORAL) software (http://www.insilico.eu/coral) in which molecular structures are represented by the simplified molecular input line entry system (SMILES) notation. Symbols inherently present in SMILES nomenclatures describe the presence of molecular fragments and therefore represent a facile approach that essentially eliminate the need to geometrically optimize molecular structures or the hassle of computing and selecting molecular descriptors. Predictive models were built in accordance with the OECD principles. Monte Carlo optimization of correlation weights of such molecular fragments provides pertinent information on structural constituents for correlating with the aromatase inhibitory activity. Results from different splits and data sub-sets indicated reliable models for predicting and interpreting the origins of aromatase inhibitory activities with the correlation coefficient (R^2) and cross-validated correlation coefficient (Q^2) in ranges of 0.6271–0.7083 and 0.6218–0.7024, respectively. Insights gained from constructed models could aid in the future design of aromatase inhibitors.

1. Introduction

Cancer is an eminent problem of public health concern as it is causes morbidity and mortality worldwide [1]. Notably, breast cancer is predominantly found in women with estrogen receptor-positive breast cancer [2]. Aromatase is an enzyme in the estrogen biosynthesis pathway that serves as the major source of estrogen production in post-manupausal women. Particularly, the catalytic activity of aromatase entails the aromatization of C19 androgens to produce C18 estrogens [3]. As high level of estrogen is associated with tumor progression, therefore, inhibition of aromatase can reduce the estrogen level thereby making it a promising target for breast cancer [2,4,5]. Aromatase inhibitors (AIs) are approved by the U.S. Food and Drug Administration as a first-line treatment for estrogen receptor-positive post-menopausal women as well as being used in cases of tamoxifen relapse [4,5]. Such AIs can be categorized as steroidal and non-steroidal on the basis of its structural chemotype whose mechanism of action is also different in which the former snugly binds in the binding pocket whereas for the latter the azole nitrogen of non-steroidal AIs coordinates to the iron-containing heme prosthetic group [6].

In efforts to reduce experimental time and cost, computational approaches are promising alternatives that can be used to gain insights on the origins of aromatase inhibitory activities. Quantitative structure–activity/property relationships (QSPR/QSAR) are robust tools for predicting the numerical data of endpoints for substances of interest that were not previously examined experimentally [7,8]. Several successful examples have been reported on the utilization of QSAR/QSPR for modeling a wide range of biological and chemical properties [9–12]. However, each QSAR model must be constructed in accordance with well-known Organisation for Economic Co-operation and Development (OECD) principles [13]. The five OECD principles are the personification of these limitations [14] which QSAR models should be accompanied by the following information: (i) a defined endpoint; (ii) an unambiguous algorithm; (iii) a defined domain of applicability; (iv) appropriate measures for goodness-of-fit, robustness and predictivity and (v) a mechanistic interpretation, if possible [15]. The widely used representation of molecular structures for QSPR/QSAR analyses is the molecular graph and/or simplified molecular input-line entry system (SMILES) [13].

* Corresponding author at: Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand. Tel.: +66 2 441 4371; fax: +66 2 441 4380.
E-mail address: chanin.nan@mahidol.ac.th (C. Nantasenamat).

http://dx.doi.org/10.1016/j.chemolab.2014.07.017
0169-7439/

A common problem in the development of QSAR/QSPR models can arise from: (i) selection of an appropriate subset of molecular descriptors from the masses of available descriptors, (ii) the vagueness of interpreting certain descriptors obtained from QSAR/QSPR modeling, and (iii) the need to geometrically optimize structures if three-dimensional descriptors are to be used. Therefore, CORAL software (available at http://www.insilico.eu/coral) is a tool that allows QSAR/QSPR analysis as a function of conformation-independent and SMILES-based descriptors while complying with the OECD principles [15–19]. CORAL software had previously been applied for modeling various biological activities and chemical properties of anti-bacterial agents [16,20], anti-cancer agents [13,18,21,22], anti-HIV agents [23], antimalarial agents [24,25], anti-neuraminidase agents [19], toxicity of compounds [15,26], degradation of pollutants [27], inhibitors of voltage-gated potassium channel subunit Kv7.2 [28], anti- human serine proteinases [29] and anticonvulsant [30] agents.

Thus, the present study employed CORAL software for constructing large-scale QSAR models for predicting the aromatase inhibitory activities of a set of 973 organic compounds (steroidal and non-steroidal AIs) based on the Monte Carlo approach. Such models afford a simple and versatile approach for discerning the origins of investigated activities directly from the SMILES notation that had been used for encoding molecular structures. The reliability of constructed QSAR models was rigorously evaluated by means of 4 subsets of data for three random splits.

2. Method

2.1. Data

The numerical data for the aromatase inhibitory activities of 973 aromatase inhibitors were taken from our previous work [6] on exploring the chemical space of aromatase inhibitors. The negative logarithmic IC_{50} values of aromatase inhibitory activity (pIC_{50}) were selected as the endpoint for QSAR analysis. Three random splits of the data into sub-training, calibration, test, and validation sets were performed. The identity of these splits was lower than 30% (Table 1). No information from the validation set was involved in building the model. In other words, compounds from the validation set are invisible in the modeling process. SMILES of compounds used in the representation of molecular structures and their distributions to sub-training, calibration, test and validation sets are provided in Supplementary Tables S1 and S2.

Table 1
Percentages of identity for random splits.

	Set	Split 1	Split 2	Split 3
Split 1	Sub-training	100.0*	34.5	32.5
	Calibration	100.0	34.8	33.6
	Test	100.0	19.2	19.4
	Validation	100.0	23.4	21.7
Split 2	Sub-training		100.0	31.9
	Calibration		100.0	27.1
	Test		100.0	24.0
	Validation		100.0	22.2
Split 3	Sub-training			100.0
	Calibration			100.0
	Test			100.0
	Validation			100.0

* $Identity(\%) = \frac{N_{i,j}}{0.5 * \left(N_i + N_j\right)} \times 100$

where $N_{i,j}$ is the number of substances distributed into the same set for both the i^{th} split and the j^{th} splits (set = sub-training, calibration, test, and validation); N_i is the number of substances distributed into the set for the i^{th} split; and N_j is the number of substances distributed into the set for the j^{th} split.

2.2. Optimal descriptor

Optimal descriptors for constructing QSAR models are based on SMILES notation as described according to the following equation [26]:

$$DCW\left(SMILES,\ Threshold,\ N_{epoch}\right) = \sum CW(S_k) + \sum CW(SS_k) + \sum CW(SSS_k) + CW(NOSP) + CW(HALO) + CW(BOND) + (CW(PAIR) \quad (1)$$

where threshold is the coefficient for classifying various molecular features extracted from SMILES into two classes: (i) active (in this case, correlation weight is involved in the modeling process) and (ii) rare (in this case, correlation weight is not involved in the modeling process). The N_{epoch} is the number of epochs using in Monte Carlo optimization giving rise to the best statistical quality for the calibration set, Sk refers to one or two symbols from SMILES (e.g. '@@', 'Cl', 'Br', etc.) that cannot be examined separately, SSk and SSSk are the combination of two and three Sk in SMILES, respectively; NOSP, HALO, BOND, and PAIR are descriptors calculated according to the presence or absence of various chemical elements and covalent bonds [26]; CW(X) is the correlation weight for a SMILES attribute (descriptor). Particularly, NOSP represents nitrogen, oxygen, sulfur and phosphorus atoms; HALO represents fluorine (F), chlorine (Cl) and bromine (Br); BOND represents double (=), triple (#) or stereochemical bonds (@ or @@); and PAIR represents the possible combination of pair atoms and/or SMILES attributes (such as, double, triple, and stereochemical bonds) that takes place in the structure together. Table 2 contains examples of the above-mentioned SMILES attributes.

Table 2
The scheme of extraction of SMILES atoms and other SMILES attributes in order to build up a model.

SMILES attribute	Examples on the representation for CORAL software
S_k	SMILES-atoms, i.e. one symbol or two symbols which cannot be examined separately, e.g. 'C' and 'Cl': this information is represented by sequences of twelve symbols: `C..........` `Cl.........`
SS_k	A combination of two SMILES-atoms 'CC' and 'CN': this information is represented by sequences of twelve symbols: `C...C.......` `C...N.......`
SSS_k	A combination of three SMILES-atoms 'CNC' and 'C#N': this information is represented by sequences of twelve symbols `C...N...C...` `C...#...N...`

BOND	The presence/absence of double ('='), triple ('#'), and stereo chemical ('@') bonds, e.g. if SMILES = "CCC(O)CC"
	`=` `#` `@` ⇨ `BOND00000000`
	`0` `0` `0`

NOSP	Presence (absence) of nitrogen, oxygen, sulfur, and phosphorus, e.g. if SMILES = "CCC(O)CC"
	`N` `O` `S` `P` ⇨ `NOSP01000000`
	`0` `1` `0` `0`

HALO	Presence (absence) of fluorine, chlorine, bromine, and iodine atoms, e.g. if SMILES = "ClCC(=O)CCl"
	`F` `Cl` `Br` ⇨ `HALO01000000`
	`0` `1` `0`

PAIR	Simultaneous presence of two SMILES-atoms from the list: F, Cl, Br, I, N, O, S, P, #, = and @; e.g. if SMILES = "ClCC(=O)CCl" the following pairs will be extracted: `++++Cl..O...` i.e. 'Cl' and 'O' `++++Cl..B2...` i.e. 'Cl' and '=' `++++O...B2...` i.e. 'O' and '='

Correlation weights are calculated by the Monte Carlo method for which they must provide the best statistical performance for the visible test set. The preferable threshold (T*) was selected using ranges of 1–5 and the preferable number of epochs (N*) was selected from 1 to 75 for searching of the best T* and N*, respectively. The preliminary computational experiments have shown that threshold larger than 5 afforded poor statistics for the model while the number of epochs larger than 75 gave no modifications to the statistics for the sub-training, calibration and test sets. Numerical data of correlation weights that afford preferable statistics for the calibration set makes it possible to calculate the endpoint value from the sub-training set as follows:

$$\text{Endpoint} = C_0 + C_1 \times \text{DCW}\left(\text{SMILES},\ \text{Threshold},\ N_{\text{epoch}}\right) \quad (2)$$

The predictive potential of the model should then be verified by means of an external validation set, which is invisible during model building. The statistical quality of the prediction (i.e. the statistical quality of the model for the test set) is a mathematical function of the threshold and the number of epochs from Monte Carlo optimization. The preferable threshold (T*) and the preferable number of epochs (N*) are parameters that provide maximal correlation coefficient between experimental and calculated values of endpoint values from the test set. Thus, roles of the four sets can be defined as follows: (i) the sub-training set is used to develop the model; (ii) calibration set is used to critique the model by checking whether the model is satisfactory for compounds that are absent from the sub-training set; (iii) the test set is used to preliminarily estimate the predictive potential of the model; and (iv) the validation set provide the final estimate on the predictive potential of the model. Further description of the software in detail is available on the Coral website (http://www.insilico.eu/coral).

2.3. Statistical assessment of QSAR models

QSAR models were evaluated by R^2 and Q^2, which corresponds to the goodness-of-fit and the goodness-of-prediction parameters, respectively. The reliability of QSAR models for interpretation was provided by the difference of R^2 and Q^2 as originally proposed by Eriksson and Johansson [31]. The Fischer (F) ratio also provide further estimate of the model's predictivity. Q^2 was calculated according to the following equation:

$$Q^2 = 1 - \frac{\sum_{i=1}^{traning} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{training} (y_i - \overline{y}_i)^2} \quad (3)$$

where y_i is the experimental value, $\hat{y}_i$ is the predicted value, and $\overline{y}_i$ is the averaged value of the entire data set and summation applies to all compounds in the sub-training set.

Y-scrambling was also performed to rule out chance correlations. This was performed as described previously by Ojha and Roy [32] in which ten probes of calculation were carried out. One probe of calculation R_r was derived as follows where **X** represents the vectors of experiment, **Y** is the vector of prediction. Firstly, exchanges of random N1 and random N2 from row X (Y is not modified) were carried out 1000 times. Secondly, $R^2_{(X,Y)}$ was then calculated from the ten aforementioned probes and denoted as R_r^2. Finally, the parameter ${}^{C}R_p^2$ was then calculated according to the following equation:

$${}^{C}R_p^2 = R \times \left(R^2 - R_r^2\right)^{1/2} \quad (4)$$

where R^2 from the non-randomized model and R_r^2 from the randomized model were utilized. ${}^{C}R_p^2$ should be greater than 0.5 for an acceptable QSAR model.

3. Results and discussion

3.1. QSAR modeling of aromatase inhibitors

In our previous study on exploring the chemical space of all known AIs [6] compiled from the literature, the derived data set was employed for classifying compounds as active and inactive by means of decision tree analysis. Compounds were represented by quantum chemical descriptors comprising of mean absolute charge, total energy, dipole moment, highest occupied molecular orbital, lowest unoccupied molecular orbital, energy gap of the HOMO and LUMO together with molecular descriptors comprising of molecular weight, rotatable bond number, number of rings, number of hydrogen bond donor, number of hydrogen bond acceptor, Ghose-Crippen octanol–water partition coefficient and topological polar surface area. The classification model provided reliable statistical quality as observed from accuracy greater than 70% for classifying active and inactive compounds.

Although extremely useful the previous model could only afford binary classification of the potential activity of compounds and the natural extension to this would be the ability to quantitatively predict the compound's aromatase inhibitory activity. Thus, this study explores the origins of aromatase inhibitory activity as a function of molecular features extracted from SMILES-based attributes. The present study also represents the first report for constructing regression QSAR models for predicting the numerical pIC_{50} values of aromatase inhibitory activities. Aromatase inhibitory activities from a large data set of 973 AIs were modeled using CORAL software. This involved the use of SMILES format for encoding the molecular description of compounds that is subsequently used for calculating (predicting) the endpoint data that is the pIC_{50} values. Such SMILES-based descriptors were then used in the construction of predictive QSAR models by means of the Monte Carlo approach. The general procedures of the QSAR modeling process are summarized in Fig. 1.

Preferable values for T* and N* to use for Monte Carlo optimization were defined in a preliminary analysis of the model calculated using threshold values in the range of 1 to 5 (Table 3) and the number of epochs ranging from 1 to 75 (Table 4). Both tables also show their respective statistical characteristics. The optimal values of T* and N* were then used for constructing the QSAR model. Results suggested that the best values of T* for splits 1, 2 and 3 were 4, 3 and 2, respectively, whereas the best values of N* were 28, 28 and 49, respectively. The data set comprising 973 AIs were divided into four groups (i.e. sub-training, calibration, test and validation sets) and each set were used for evaluating the predictive performance as shown below:

Split 1:

$$\begin{aligned}
&pIC_{50} = 0.0000\ (\pm 0.01454) + 0.0772\ (\pm 0.0002) \times \text{DCW(SMILES, 4, 28)}\\
&n = 320, R^2 = 0.6152, Q^2 = 0.6105, R^2 - Q^2 = 0.0047, s = 0.803, F = 508\ \text{(sub−training set)}\\
&n = 323, R^2 = 0.6152, Q^2 = 0.6102, R^2 - Q^2 = 0.0050, s = 0.874\ \text{(calibration set)}\\
&n = 174, R^2 = 0.6194, Q^2 = 0.6101, R^2 - Q^2 = 0.0093, s = 0.740\ \text{(test set)}\\
&n = 156, R^2 = 0.6907, Q^2 = 0.6826, R^2 - Q^2 = 0.0081, s = 0.615\ \text{(validation set)}
\end{aligned} \quad (5)$$

Split 2:

$$\begin{aligned}
&pIC_{50} = 0.0028\ (\pm 0.0155) + 0.0797\ (\pm 0.0002) \times \text{DCW(SMILES, 3, 28)}\\
&n = 300, R^2 = 0.6495, Q^2 = 0.6446, R^2 - Q^2 = 0.0049, s = 0.799, F = 552\ \text{(sub−training set)}\\
&n = 281, R^2 = 0.6495, Q^2 = 0.6445, R^2 - Q^2 = 0.0050, s = 0.835\ \text{(calibration set)}\\
&n = 180, R^2 = 0.6739, Q^2 = 0.6663, R^2 - Q^2 = 0.0076, s = 0.662\ \text{(test set)}\\
&n = 212, R^2 = 0.6696, Q^2 = 0.6633, R^2 - Q^2 = 0.0063, s = 0.6593\\
&\text{(validation set)}
\end{aligned}$$

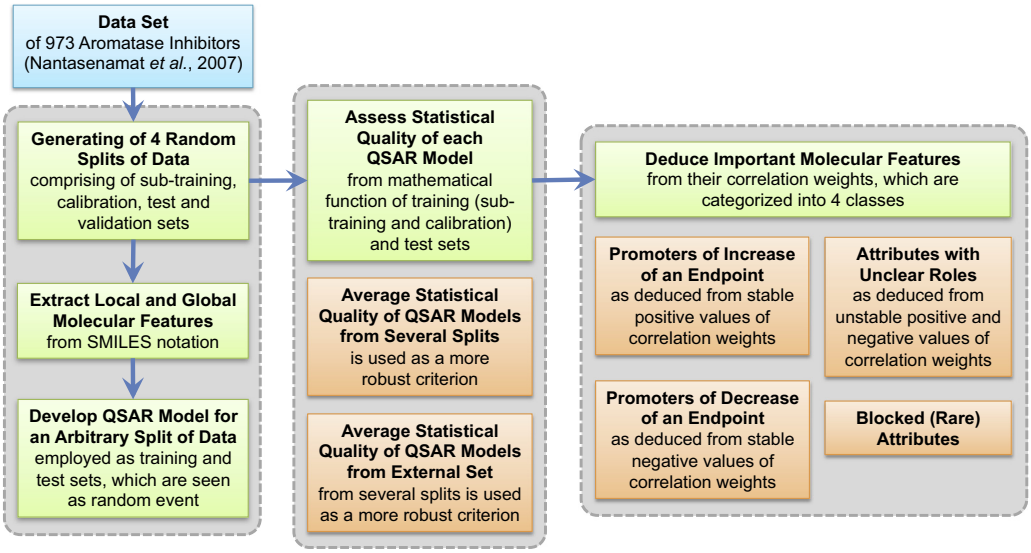

Fig. 1. Schematic illustration on the workflow of the predictive QSAR modeling performed herein.

Split 3:

$$\begin{aligned}
&\mathrm{pIC_{50}} = 0.0001\ (\pm 0.01509) + 0.1063\ (\pm 0.0003) \times \mathrm{DCW(SMILES,\ 2, 49)} \\
&n = 314, R^2 = 0.6271, Q^2 = 0.6224, R^2 - Q^2 = 0.0047, s = 0.794, F = 525\ \text{(sub−training set)} \\
&n = 279, R^2 = 0.6273, Q^2 = 0.6218, R^2 - Q^2 = 0.0055, s = 0.864\ \text{(calibration set)} \\
&n = 186, R^2 = 0.7083, Q^2 = 0.7024, R^2 - Q^2 = 0.0059, s = 0.666\ \text{(test set)} \\
&n = 194, R^2 = 0.6568, Q^2 = 0.6499, R^2 - Q^2 = 0.0069, s = 0.653\ \text{(validation set)}
\end{aligned} \quad (7)$$

For Eqs. (5)–(7), n is the number of compounds in each set, R^2 is the squared correlation coefficient value, Q^2 is the leave-one-out cross-validation coefficient of determination, $R^2 - Q^2$ is the difference of R^2 and Q^2, s is standard error of estimation and F is the Fischer ratio. In addition, DCW from Eqs. (5)–(7) displayed the best values of T* and N* (as shown in Tables 3 and 4) for constructing QSAR models and providing maximum correlation between the experimental and predicted ($\mathrm{pIC_{50}}$) values. For example, DCW (SMILES, 2, 24) from split 1 means that T* = 2 and N* = 4.

It can be seen that the obtained QSAR equations afforded reliable statistical quality for sub-training, calibration, test and validation sets according to criterion described by Tropsha et al. [33] for QSAR models where $R^2 > 0.6$ and $Q^2 > 0.5$ indicates predictive models. Plots of experimental versus predicted $\mathrm{pIC_{50}}$ values from models calculated with Eqs. (5)–(7) are presented in Fig. 2. It can thus be seen that the predicted $\mathrm{pIC_{50}}$ values of compounds were in good correlation with its experimental values. The numerical data of experimental and predicted $\mathrm{pIC_{50}}$ values with DWC from splits 1 to 3 as calculated from Eqs. (5)–(7) are provided in Supplementary Tables S1 and S2.

Furthermore, the reliability of constructed models was also assessed from their $R^2 - Q^2$ values, which is a metric that accounts for the fraction of **Y**-data explained by accumulated chance correlations where values greater than 0.2–0.3 is indicative of the risk for chance correlations or the presence of outliers in the data set. It was observed that all four sets from three splits of QSAR modeling provided extremely low $R^2 - Q^2$ values in the range of 0.0045 and 0.0091 that is well below the criterion and thereby confirming the reliability of constructed models for further interpretations. Concomitant with this result, further test of chance correlations was evaluated by Y-scrambling in which the **Y** value (i.e. $\mathrm{pIC_{50}}$) is shuffled or randomly reordered with respect to its associated **X** descriptors. Thus, 1000 trials of Y-scrambling were performed in ten separate runs for all three splits and the average value for each run is shown in Table 5. Results from Y-scrambling verify the predictivity of constructed models with R^2 values <0.0517.

Table 3
Definition of the preferable threshold (T*). Optimal threshold is indicated in bold.

Split	Threshold	Correlation coefficient (R^2)					Preferable
		Probe 1	Probe 2	Probe 3	Average	Dispersion	
1	1	0.5857	0.6129	0.5839	0.5942	0.0133	T* = 4
	2	0.6210	0.6008	0.6155	0.6124	0.0085	
	3	0.6199	0.6071	0.5838	0.6036	0.0149	
	4	0.6463	0.6186	0.6193	**0.6281**	0.0129	
	5	0.5832	0.5972	0.6062	0.5955	0.0095	
2	1	0.6719	0.6750	0.6724	0.6731	0.0014	T* = 3
	2	0.6643	0.6593	0.6742	0.6660	0.0062	
	3	0.6766	0.6889	0.6695	**0.6783**	0.0080	
	4	0.6612	0.6492	0.6634	0.6580	0.0063	
	5	0.6414	0.6351	0.6366	0.6377	0.0027	
3	1	0.6893	0.6928	0.6938	0.6920	0.0019	T* = 2
	2	0.7104	0.7136	0.7112	**0.7118**	0.0013	
	3	0.6971	0.6965	0.7034	0.6990	0.0031	
	4	0.6904	0.6838	0.6849	0.6864	0.0029	
	5	0.6604	0.6576	0.6601	0.6593	0.0013	

Table 4
Definition of the number of epochs of the Monte Carlo optimization (N*). Optimal number of epochs is indicated in bold.

Split	Threshold	Number of epochs					Preferable
		Probe 1	Probe 2	Probe 3	Average	Dispersion	
1	1	15	14	27	18.67	5.91	N* = 28
	2	22	32	17	23.67	6.24	
	3	31	14	22	22.33	6.94	
	4	28	35	21	**28.00 ≈ 28**	5.72	
	5	20	35	35	30.00	7.07	
2	1	7	13	11	10.33	2.49	N* = 28
	2	33	14	12	19.67	9.46	
	3	16	35	32	**27.67 ≈ 28**	8.34	
	4	49	41	28	39.33	8.65	
	5	11	34	26	23.67	9.53	
3	1	47	44	48	46.33	1.70	N* = 49
	2	49	49	48	48.67	0.47	
	3	46	50	50	**48.67 ≈ 49**	1.89	
	4	50	50	49	49.67	0.47	
	5	49	50	50	49.67	0.47	

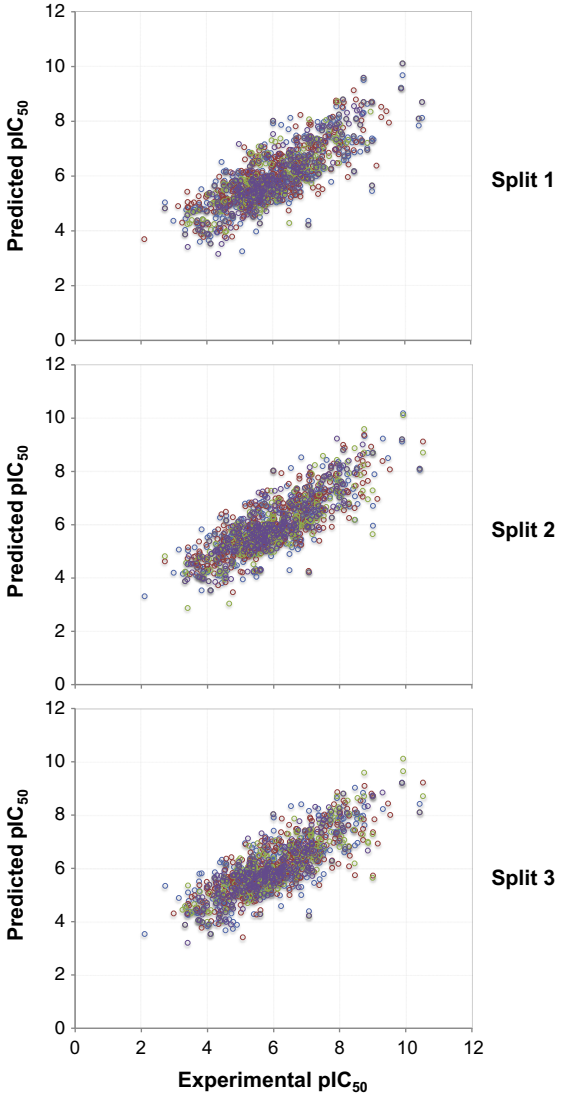

Fig. 2. Plot of experimental versus predicted pIC_{50} values as obtained from 3 splits of QSAR modeling. Calibration, sub-training, test and validation sets are shown in blue, red, green and purple, respectively.

In assessing the general level of performance of SMILES-based QSAR models presented herein, it is necessary to consider the results of previously reported QSAR models. The first QSAR regression model on a diverse set of aromatase inhibitors was reported by Roy and Roy [34] where they employed both 2D and 3D descriptors to afford the following statistical performance: $n = 87$ and $R^2 = 0.691$ for training set while $n = 29$ and $Q^2 = 0.63$ for the test set. A subsequently reported QSAR model on a set of flavonoid derivatives from Narayana et al. [35] was developed using 3D descriptors resulting in the following statistical performance: $n = 39$ and $R^2 = 0.825$ for the training set while $n = 18$ and $Q^2 = 0.815$ for the test set. Recently, we reported a QSAR model on a set of 1,2,3-triazole analogs of letrozole [36] with statistical performance as follows: $n = 40$ and $R^2 = 0.7719$ for the training set while $n = 40$ and $Q^2 = 0.6932$ for the leave-one-out cross-validated test set. Moreover, we also reported a QSAR model on two sets of flavonoid derivatives [37] and obtained the following results for the first set: $n = 33$ and $R^2 = 0.9910$ for the training set while $n = 33$ and $Q^2 = 0.9736$ for the leave-one-out cross-validated test set. The second set produced the following results: $n = 19$ and $R^2 = 0.9536$ for the training set while $n = 19$ and $Q^2 = 0.8316$. Comparison of the statistical quality of QSAR models suggested in this work with the aforementioned models indicated that CORAL software gave reasonably good model for aromatase inhibitory activity.

Table 5
Y-scrambling performed for 1000 trials in ten separate runs. The average R^2 values for each run are shown.

Split 1			
	Training	Calibration	Test
	320	323	174
Original	0.6152	0.6152	0.6194
1	0.0080	0.0007	0.0175
2	0.0051	0.0053	0.0091
3	0.0090	0.0019	0.0082
4	0.0008	0.0046	0.0138
5	0.0018	0.0138	0.0003
6	0.0084	0.0057	0.0009
7	0.0010	0.0002	0.0005
8	0.0091	0.0025	0.0083
9	0.0004	0.0260	0.0015
10	0.0021	0.0265	0.0005
R_r^2 [a]	0.0046	0.0087	0.0060
${}^{c}R_p^2$ [b]	0.6129	0.6108	0.6164
Split 2			
	Training	Calibration	Test
	300	281	180
Original	0.6495	0.6495	0.6739
1	0.0097	0.0040	0.0006
2	0.0108	0.0155	0.0171
3	0.0132	0.0010	0.0010
4	0.0001	0.0153	0.0090
5	0.0001	0.0118	0.0299
6	0.0079	0.0273	0.0283
7	0.0022	0.0057	0.0020
8	0.0111	0.0001	0.0360
9	0.0016	0.0129	0.0203
10	0.0007	0.0018	0.0252
R_r^2 [a]	0.0057	0.0095	0.0170
${}^{c}R_p^2$ [b]	0.6466	0.6447	0.6654
Split 3			
	Training	Calibration	Test
	314	279	186
Original	0.6271	0.6273	0.7083
1	0.0218	0.0073	0.0303
2	0.0027	0.0207	0.0006
3	0.0014	0.0006	0.0224
4	0.0091	0.0060	0.0017
5	0.0046	0.0117	0.0048
6	0.0043	0.0001	0.0276
7	0.0023	0.0001	0.0139
8	0.0029	0.0012	0.0001
9	0.0026	0.0038	0.0050
10	0.0099	0.0001	0.0033
R_r^2 [a]	0.0062	0.0052	0.0110
${}^{c}R_p^2$ [b]	0.6240	0.6247	0.7028

[a] Average randomized R^2.
[b] ${}^{c}R_p^2 = R \times (R^2 - R_r^2)^{1/2}$ where ${}^{c}R_p^2$ should be greater than 0.5*.
* Please refer to [32] for further details.

3.2. Interpretation of structure–activity relationship

The SMILES molecular fragments were interpreted for exploring chemical information that is involved in aromatase inhibitory activity via the analysis of correlation weights obtained from QSAR modeling. This was performed by dividing the data samples into the following four classes: (i) list of promoters of pIC_{50} increase (all correlation weights are positive); (ii) list of promoters of pIC_{50} decrease (all correlation weights are negative); (iii) attributes with unclear role (there are

both negative and positive correlation weights); and (iv) blocked (rare) attributes. Lists of the most significant promoters of activity are displayed in Supplementary Table S3 that considers the ten top-ranking fragments for increasing and decreasing the activity. Top-ranking fragments for increasing the activity are (i) the presence of cyclic rings in the molecular structure (e.g. attributes of "1....." and "2....."); (ii) the absence of halogens (HALO00000000); (iii) the presence of double bond (BOND10000000); and (iv) the presence in the molecular structure of oxygen atoms together with double bonds that are disconnected in the structure (++++O–B2==). Furthermore, top-ranking fragments for decreasing the activity are (i) branching (i.e. presence of brackets in the SMILES); (ii) the presence of nitrogen and double bonds that are disconnected in the structure (++++N–B2==); and (iii) the presence of oxygen atoms connected via double bonds ("=…O…(…"). Thus, the approach employed herein provides mechanistic interpretations for deducing how molecular fragments may exert its influence on the aromatase inhibitory activity in increasing or decreasing the activity.

4. Conclusions

In this study, CORAL software was employed for constructing QSAR models for predicting the aromatase inhibitory activities of a large number of aromatase inhibitors. CORAL software gives reliable predictive models for aromatase inhibitory activities using SMILES-based descriptors, which are used to derive the correlation weights for molecular features using Monte Carlo method. The predictive performance was validated using different splits of the data set (i.e. sub-training, calibration, test and validation sets) for constructing QSAR models in accordance with OECD guidelines. Predictive QSAR models provided herein offers the potential to design novel aromatase inhibitors for estrogen receptor-positive breast cancer.

Conflict of interest

The authors declare that there are no know conflict of interest.

Acknowledgements

This research is supported by the Goal-Oriented Research Grant from Mahidol University (B.E. 2555-2557)to C.N. Additionally, A.P.T. and A.A.T. acknowledge support from the EC project NANOPUZZLES (Project Reference: 309837), EU FP7 project PreNanoTox (Contract Number: 309666) and EC project CALEIDOS (Project Number: LIFE11-INV/IT 00295).

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.chemolab.2014.07.017.

References

[1] A. Jemal, F. Bray, M.M. Center, J. Ferlay, E. Ward, D. Forman, Global cancer statistics, CA Cancer J. Clin. 61 (2011) 69–90.
[2] A.D. Favia, O. Nicolotti, A. Stefanachi, F. Leonetti, A. Carotti, Computational methods for the design of potent aromatase inhibitors, Expert Opin. Drug Discov. 8 (2013) 395–409.
[3] N. Suvannang, C. Nantasenamat, C. Isarankura-Na-Ayudhya, V. Prachayasittikul, Molecular docking of aromatase inhibitors, Molecules 16 (2011) 3597–3617.
[4] J. Narashimamurthy, A.R. Rao, G.N. Sastry, Aromatase inhibitors: a new paradigm in breast cancer treatment, Curr. Med. Chem. Anticancer Agents 4 (2004) 523–534.
[5] J.K. Litton, B.K. Arun, P.H. Brown, G.N. Hortobagyi, Aromatase inhibitors and breast cancer prevention, Expert. Opin. Pharmacother. 13 (2012) 325–331.
[6] C. Nantasenamat, H. Li, P. Mandi, A. Worachartcheewan, T. Monnor, C. Isarankura-Na-Ayudhya, V. Prachayasittikul, Exploring the chemical space of aromatase inhibitors, Mol. Divers. 17 (2013) 661–677.
[7] C. Nantasenamat, C. Isarankura-Na-Ayudhya, V. Prachayasittikul, Advances in computational methods to predict the biological activity of compounds, Expert Opin. Drug Discov. 5 (2010) 633–654.
[8] C. Nantasenamat, C. Isarankura-Na-Ayudhya, T. Naenna, V. Prachayasittikul, A practical overview of quantitative structure–activity relationship, EXCLI J 8 (2009) 74–88.
[9] M. Khoshneviszadeh, N. Edraki, R. Miri, A. Foroumadi, B. Hemmateenejad, QSAR study of 4-aryl-4H-chromenes as a new series of apoptosis inducers using different chemometric tools, Chem. Biol. Drug Des. 79 (2012) 442–458.
[10] Y. Uesawa, K. Mohri, M. Kawase, M. Ishihara, H. Sakagami, Quantitative structure-activity relationship (QSAR) analysis of tumor-specificity of 1,2,3,4-tetrahydroisoquinoline derivatives, Anticancer Res. 31 (2011) 4231–4238.
[11] A. Worachartcheewan, C. Nantasenamat, C. Isarankura-Na-Ayudhya, S. Prachayasittikul, V. Prachayasittikul, Predicting the free radical scavenging activity of curcumin derivatives, Chemometr. Intell. Lab. Syst. 109 (2011) 207–216.
[12] 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*, Chem. Pap. 67 (2013) 1462–1473.
[13] A.A. Toropov, A.P. Toropova, E. Benfenati, Additive SMILES-based carcinogenicity models: probabilistic principles in the search for robust predictions, Int. J. Mol. Sci. 10 (2009) 3106–3127.
[14] Organisation for Economic Co-operation and Development, Guidance document on the validation of (quantitative) structure-activity relationship [(Q)SAR] models, http://www.oecd.org/dataoecd/55/35/38130292.pdf (2007).
[15] A.P. Toropova, A.A. Toropov, A. Lombardo, A. Roncaglioni, E. Benfenati, G. Gini, Coral: QSAR models for acute toxicity in fathead minnow (*Pimephales promelas*), J. Comput. Chem. 33 (2012) 1218–1223.
[16] A.A. Toropov, A.P. Toropova, E. Benfenati, G. Gini, D. Leszczynska, J. Leszczynski, G. De Nucci, QSAR models for inhibitors of physiological impact of *Escherichia coli* that leads to diarrhea, Biochem. Biophys. Res. Commun. 432 (2013) 214–225.
[17] A.A. Toropov, A.P. Toropova, E. Benfenati, G. Gini, D. Leszczynska, J. Leszczynski, CORAL: QSPR model of water solubility based on local and global SMILES attributes, Chemosphere 90 (2013) 877–880.
[18] A.P. Toropova, A.A. Toropov, CORAL software: prediction of carcinogenicity of drugs by means of the Monte Carlo method, Eur. J. Pharm. Sci. 52 (2014) 21–25.
[19] A. Worachartcheewan, C. Nantasenamat, C. Isarankura-Na-Ayudhya, V. Prachayasittikul, QSAR study of H1N1 neuraminidase inhibitors from influenza A virus, Lett. Drug Des. Discov. 11 (2014) 420–427.
[20] A.A. Toropov, A.P. Toropova, E. Benfenati, G. Gini, T. Puzyn, D. Leszczynska, J. Leszczynski, Novel application of the CORAL software to model cytotoxicity of metal oxide nanoparticles to bacteria *Escherichia coli*, Chemosphere 89 (2012) 1098–1102.
[21] A.A. Toropov, A.P. Toropova, E. Benfenati, G. Gini, D. Leszczynska, J. Leszczynski, CORAL: classification model for predictions of anti-sarcoma activity, Curr. Top. Med. Chem. 12 (2012) 2741–2744.
[22] L.M.A. Mullen, P.R. Duchowicz, E.A. Castro, QSAR treatment on a new class of triphenylmethyl-containing compounds as potent anticancer agents, Chemometr. Intell. Lab. Syst. 107 (2011) 269–275.
[23] A.P. Toropova, A.A. Toropov, J.B. Veselinovic, F.N. Miljkovic, A.M. Veselinovic, QSAR models for HEPT derivates as NNRTI inhibitors based on Monte Carlo method, Eur. J. Med. Chem. 77 (2014) 298–305.
[24] V.H. Masand, A.A. Toropov, A.P. Toropova, D.T. Mahajan, QSAR models for anti-malarial activity of 4-aminoquinolines, Curr. Comput. Aided Drug Des. 10 (2014) 75–82.
[25] E. Ibezim, P.R. Duchowicz, E.V. Ortiz, E.A. Castro, QSAR on aryl-piperazine derivatives with activity on malaria, Chemometr. Intell. Lab. Syst. 110 (2012) 81–88.
[26] A.P. Toropova, A.A. Toropov, E. Benfenati, G. Gini, D. Leszczynska, J. Leszczynski, CORAL: quantitative structure–activity relationship models for estimating toxicity of organic compounds in rats, J. Comput. Chem. 32 (2011) 2727–2733.
[27] A.A. Toropov, A.P. Toropova, B.F. Rasulev, E. Benfenati, G. Gini, D. Leszczynska, J. Leszczynski, CORAL: QSPR modeling of rate constants of reactions between organic aromatic pollutants and hydroxyl radical, J. Comput. Chem. 33 (2012) 1902–1906.
[28] P.G. Achary, Simplified molecular input line entry system-based optimal descriptors: QSAR modelling for voltage-gated potassium channel subunit Kv7.2, SAR QSAR Environ. Res. 25 (2014) 73–90.
[29] J. García, P.R. Duchowicz, M.F. Rozas, J.A. Caram, M.V. Mirífico, F.M. Fernández, E.A. Castro, A comparative QSAR on 1,2,5-thiadiazolidin-3-one 1,1-dioxide compounds as selective inhibitors of human serine proteinases, J. Mol. Graph. Model. 31 (2011) 10–19.
[30] J.C.G. Martinez, P.R. Duchowicz, M.R. Estrada, G.N. Zamarbide, E.A. Castro, QSAR study and molecular design of open-chain enaminones as anticonvulsant agents, Int. J. Mol. Sci. 12 (2011) 9354–9368.
[31] L. Eriksson, E. Johansson, Multivariate design and modeling in QSAR, Chemometr. Intell. Lab. Syst. 34 (1996) 1–19.
[32] P.K. Ojha, K. Roy, Comparative QSARs for antimalarial endochins: importance of descriptor-thinning and noise reduction prior to feature selection, Chemometr. Intell. Lab. Syst. 109 (2011) 146–161.
[33] A. Tropsha, P. Gramatica, V.K. Gombar, The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models, QSAR Comb. Sci. 22 (2003) 69–77.
[34] P.P. Roy, K. Roy, Docking and 3D-QSAR studies of diverse classes of human aromatase (CYP19) inhibitors, J. Mol. Model. 16 (2010) 1597–1616.
[35] B.L. Narayana, D. Pran Kishore, C. Balakumar, K.V. Rao, R. Kaur, A.R. Rao, J.N. Murthy, M. Ravikumar, Molecular modeling evaluation of non-steroidal aromatase inhibitors, Chem. Biol. Drug Des. 79 (2012) 674–682.

[36] 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, Eur. J. Med. Chem. 69 (2013) 99–114.

[37] 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, Chem. Pap. 68 (2014) 697–713.