Original Article

TIPS Trends in Pharmaceutical Sciences 2018: 4(4): 205-218.

# QSAR analysis of some azole derivatives as potent aromatase inhibitors

Razeih Sabet[1], Soghra Khabnadideh[1,2,*], Samaneh Mohseni[1], Ayyub Mojaddami[3], Zahra Rezaei[1,*]

[1]Department of Medicinal Chemistry, School of Pharmacy, Shiraz University of Medical Sciences, Shiraz, Iran.
[2]Pharmaceutical Sciences Research Center, Shiraz University of Medical Sciences, Shiraz, Iran.
[3]Faculty of Pharmacy, Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran.

## Abstract

A high proportion of breast tumors are hormone-dependent, implying that endogenous estrogens play a critical role in cancer cell proliferation. One of the most effective strategies for the treatment of breast cancer is the reduction of estrogen level by inhibiting aromatase enzyme, which is responsible for catalyzing the rate-limiting step in estrogen biosynthesis. A series of azole derivatives as potential aromatase inhibitors were subjected to two different drug design methodologies: QSAR and molecular docking simulation. MLR, FA-MLR, PCR, and GA-PLS were employed to explore connections between the structural parameters and aromatase inhibitory activity. GA-PLS represented superior results and a model with a high statistical quality ($R^2$=0.86 and $Q^2$=0.83) for predicting the inhibitory activity. The results can provide useful information for the development of more potent aromatase inhibitors.

*Keywords:* Aromatase inhibitor, Azole derivatives, QSAR.

## 1. Introduction

The two main categories of drug design methods, namely (i) ligand-based and (ii) structure-based approaches, are usually regarded as complementary methodologies in modern computational drug designing. These approaches rely on the physicochemical and structural information of the drug molecule, and the three dimensional structure of the target macromolecule in its binding (active) site, respectively.

Quantitative structure activity relationships (QSAR) studies are the most common researches in the ligand-based methods (1,2). In QSAR researches, a statistical model correlating the structure and biological activity is found. This model is then applied to screen any number of molecules even those that have not yet been synthesized to predict their biological activity. Although numerous physicochemical descriptors have been used in such studies, QSAR models are unable to explain some facts in ligand-target interactions. The strength of hydrogen bonds, the influence of desolation energies on drug-receptor bindings, and steric interactions of the ligand with the binding site are examples of limitations of QSAR studies.

The considerable body of information about the high resolution three dimensional structures of drug targets, including receptors, channels, enzymes, and transporter proteins, provides a growing basis for the structure-based drug design approach. Ligand-target interactions are modeled through molecular docking simulation technique (3-6).The techniques and algorithms employed for ligand-protein binding energy prediction have a number of limitations. A time-consuming process for large numbers of molecules, insufficient sampling of protein flexibility, and inaccurate scoring functions applied to rank the ligands interacting

*Corresponding Author*: Zahra Rezaei & Soghra Khabnadideh, Department of Medicinal Chemistry, School of Pharmacy, Shiraz University of Medical Sciences, Shiraz, Iran.
Email: rezaeiza@sums.ac.ir; khabns@sums.ac.ir


Recieved: 20/10/2018; Accepted: 09/11/2018

with the target protein are some of these limitations. Hence, the simultaneous application of the two approaches, ligand-based and structure-based, helps medicinal chemists design more potent ligands.

Aromatase is a cytochrome P450 enzyme complex, which catalyzes the conversion of androgens to estrogens. Inhibition of this enzyme is crucial for the treatment of estrogen-dependent breast cancer (7-12). Aromatase inhibitors are used as the first-line therapy in postmenopausal women with metastatic; hormone receptor-positive breast cancer (9-10). Aromatase inhibitors are used in some non-tumorigenic conditions such as precocious puberty and gynecomastia in males (9). Aromatase inhibitors include both steroidal and non-steroidal subtypes (7). Some aromatase inhibitors with steroidal structure such as exemestane are enzyme activated irreversible inhibitors of aromatase. Non-steroidal aromatase inhibitors with azole moiety such as letrozole, fadrozole, and anstrazole are competitive inhibitors and bind to the $Fe^{2+}$ present in heme group of the enzyme (7-13). Other non-steroidal aromatase inhibitors have the flavon structure such as chrysin that is a natural product and has an in vitro inhibitory action similar to aminoglutethimide. Chrysin is not used therapeutically as an aromatase inhibitor due to its low oral bioavailability (13).

In the present paper, two different drug design methodologies, QSAR and molecular docking simulations, were applied for a series of 200 azole analogues with the ability to inhibit the aromatase enzyme (14). A large descriptor set, including topological, geometrical, constitutional, functional group, electrostatic, and chemical factors, was used to describe the physicochemical properties of the molecules. Different statistical methods were applied to model the relationship between the structural features and the aromatase inhibitory activity of the studied compounds. These methods were: (i) multiple linear regression (MLR) (ii) genetic algorithm-partial least squares (GA-PLS), (iii) MLR with factor analysis as the data pre-processing step for variable selection (FA-MLR), and (iv) principal component regression (PCR).

## 2. Material and methods

### *2.1. Equipment*

Two-dimensional (2D) structures of molecules were drawn using Hyperchem 7.0 software. The optimized geometries were obtained with semi-empirical AM1 Hamiltonian in the Hyperchem program using the Polak-Ribiere algorithm until the root mean square gradient was 0.01 kcal. mol-1. The resulted geometry was transferred into Dragon program package developed by Milano Chemometrics and QSAR Group (15). MATLAB software (version 7.1 Math Work Inc.) was used for model generation and validation of methods.

### *2.2. Data set and descriptor generation*

The biological data used in this study were the inhibitory activity (in terms of –logIC50) of a set of azole derivatives (16-31). These structures were then used for generating molecular descriptors as independent variables. The large numbers of molecular descriptors were calculated using Hyperchem and Dragon package. Some chemical parameters including molecular volume (MV), molecular surface area (MSA), hydrophobicity (LogP), hydration energy (HE), and molecular polarizability (MP) were calculated using Hyperchem Software. Different constitutional, topological, geometrical, and functional group descriptors were extracted with Dragon software for each molecule.

### *2.3. Data pre-processing*

In order to test the developed model performances, 30 % of the molecules (60 out of 200) were selected as the test set molecules. The Kennard and Stones algorithm for splitting datasets into training and test subsets was exploited for this purpose. All the calculated descriptors were collected in a data matrix D with a dimension of (n×k), where n is the number of molecules and k is the number of descriptors, respectively. In each group, the calculated descriptors were searched for constant or near-constant values for all molecules, and those detected were removed from the original data matrix. The correlation of descriptors with each other and with the activity data was determined. Then data matrix containing the total descriptors was subjected to principal component

analysis, and the first two principal components were plotted against each other (Figure 1A). The outlier data were assigned and deleted from the data set giving a total of 200 aromatase inhibitors.

### *2.4. Data screening and model building*

The selected descriptors from each class and the experimental data were analyzed by the stepwise regression using SPSS software (version 18.0). The calculated descriptors were collected in a data matrix whose number of rows and columns were the number of molecules and descriptors, respectively. MLR and GA-PLS, FA-MLR, and PCRA methods were used to derive the QSAR equations. The resulted models were validated by leave-one out cross-validation procedure (using MATLAB software) to check their predictability and robustness. However, this procedure did not produce good results, therefore, GA-PLS was used to select the best variables.

Application of PLS allows the construction of larger QSAR equations, while still avoiding over-fitting and eliminating most variables. PLS is normally used in combination with cross-validation to obtain the optimum number of components (32-34). The PLS regression method used in this study was the NIPALS-based algorithm existed in the chemometrics toolbox of MATLAB software (version 7.1 Math work Inc.). Leave-one-out cross-validation procedure was used to obtain the optimum number of factors based on the Haaland and Thomas F-ratio criterion (35). FA-MLR was also performed on the dataset. Factor analysis (FA) was used to reduce the number of variables and to detect structure in the relationships between them. This data-processing step is applied to identify the important predictor variables and to avoid collinearities among them (36). PCRA was also tried for the dataset along with FA-MLR. With PCRA, collinearities among X variables are not a disturbing factor and the number of variables included in the analysis may exceed the number of observations (37). In this method, factor scores, as obtained from FA, are used as the predictor variables (36). In PCRA, all descriptors are assumed to be important while the aim of factor analysis is to identify relevant descriptors.

### *2.5. Variable importance in the projection (VIP)*

In order to investigate the relative importance of the variable appeared in the final model obtained by GA-PLS method, VIP was employed (38). VIP values reflect the importance of terms in the PLS model. According to Erikson et al. X-variables (predictor variables) could be classified according to their relevance in explaining y (predicted variable), so that VIP>1.0 and VIP<0.8 mean highly or less influential, respectively, and 0.8<VIP<1.0 means moderately influential (38).

## 3. Results and Discussion

### *3.1. MLR analysis*

In the first step, separate stepwise selection-based MLR analyses were performed using different types of descriptors. Then, an MLR equation was obtained utilizing the pool of all calculated descriptors. The results are summarized in Table 1.

Correlation coefficient ($R^2$) matrix for the descriptors used in different MLR equations is shown in Table 2.

Collinear descriptors degrade the performance of MLR equations, and such models have lowered the prediction ability. As shown in Table 1, the QSAR models obtained for different derivatives using different sets of molecular descriptors are listed. Table 1 provides the resulted equations for the studied compounds. The first equation of Table 1 was found by using chemical descriptors ($E_1$). This equation explained the negative effect of hydration energy and partition coefficient of molecules on the aromatase inhibitory activity. Equation $E_2$ shows that among constitutional descriptors, mean atomic Vander Waals volume (MV) and number of nitrogen atoms (nN) have a positive effect on aromatase inhibitory activity. Number of chlorine atoms (nCl) has a negative effect on inhibitory activity. When the number of chlorine decreases, the partition coefficient also decreases; thus, the inhibitory activity increases. The presence of 9-membered rings (NR09) decreased inhibitory activity. Equation $E_3$ of Table 1 demonstrates the effect of topological descriptors. It includes the negative effects of sum of topological distances between O..O (T(O..O)), Randic-type eigenvector-based index from adjacency matrix (VRA1) and

Kier symmetry index (S0K) and the positive effect of structural information content (SIC3), sum of topological distances between N..N (T(N..N)), ratio of multiple path count to path counts (PCR), mean information index on atomic composition (ACC), and mean information content on the distance degree equality (IDDE) on aromatase inhibitory activity. The MLR equation obtained from the pool of geometrical descriptors ($E_4$) explained the positive effect of sum of geometrical distances between N..N (G(N..N)), length-to-breadth ratio by WHIM (L/BW), and sum of geometrical distances between N..O(G(N..O)) and the negative effect of 3D-Balaban index (J3O), sum of geometrical distances between O..O (G(O..O)), and span R (SPAN) of aromatase inhibitory activity. The effect of functional groups on aromatase inhibitory activity of the studied compounds has been described by equation $E_5$ in Table 1. The negative signs of NCS, NCRHR, n=CHR, and NRORPH indicate that molecules with lower number of thioketones, number of ring tertiary C(sp3), number of secondary C(sp2), and number of ethers (aromatic) bind more strongly to aromatase. On the other hand, the number of nitriles aliphatic (nCN), number of tertiary amines aliphatic (nNR2), number of phenols (nOHPH), number of sulfurs (nRSR), and nN-NPh: number of N hydrazines (aromatic) represent direct effect on the inhibitory activity of the compounds.

The equation obtained from the effect of charge parameter on aromatase inhibitory activity of the studied compounds has been described as $E_6$ in Table 1. The negative coefficient of PCWTe indicates that partial charge weighted topological electronic charge is not favorable for binding affinity. Equation $E_7$ in Table 1 demonstrates the effect of Mol-walk descriptors. This two-parametric equation does not have a high statistical quality, which suggests that the aromatase inhibitory activity of the studied molecules is not highly dependent on the type of Mol-walk group; but it is dependent on the self-returning walk count of order 05 (SRW05), molecular walk count of order 08 (MWC08). The BCUT equation ($E_8$) shows the importance of BCUT factors on aromatase inhibitory activity. Equations $E_9$ and $E_{10}$ in Table 1 demonstrate the effect of Galves and 2D descriptors. These four-parametric equations do not have a high statistical quality, which suggests that the aromatase inhibitory activity of the studied molecules is not highly dependent on the type of Galves and 2D descriptors. The effects of RDF and 3D groups on aromatase inhibitory activity of the studied compounds have been described by equations $E_{11}$ and $E_{12}$ in Table 1. Equation $E_{13}$ in Table 1 demonstrates the effect of WHIM descriptors. This tree-parametric equation does not have a high statistical quality, which suggests that the aromatase inhibitory activity of the studied molecules is not highly dependent on WHIM descriptors. $E_{14}$ explained the positive and negative effects of Getaway descriptors on inhibitory activity. The MLR equation obtained from Atom-center descriptors ($E_{15}$) has a high statistical quality explained the positive and negative effects of these descriptors. This equation describes the structure-activity relationships better than those obtained from the equations $E_1$- $E_{14}$.

The last Equation ($E_{16}$) was obtained from a collection of all descriptors. Stepwise selection and elimination of variables produced a thirteen-parametric QSAR equation. This equation shows that constitutional (MV, nN, nCl), functional (nNR2, nOHPH, nCrHR), BCUT (BELV1), topological (IDDE, VRA1), 2D (MATS4e), geometrical (SPAN), RDF020V (RDF), and Getaway (R3p+), SED (SNQ8) parameters are major factors affecting the aromatase inhibitory activity of the compounds. Among these descriptors, MV, nN, nNR2, nOHPH, IDDE, and R3p+ have positive effects, while the other descriptors have negative effects on the aromatase inhibitory activity.

### *3.2. GA-PLS analysis*

The values of experimental pIC50 were 3.6-9.8 and the predicted $IC_{50}$ were 4-9 for MLR and FA-MLR methods and 3-9 for GA-PLS and PCR methods. The predicted $IC_{50}$ were close to the experimental $IC_{50}$. In PLS analysis, the descriptors data matrix is decomposed to orthogonal matrices with an inner relationship between the dependent and independent variables. Therefore, unlike MLR analysis, the multicolinearity problem in the descriptors is omitted by PLS analysis. Because a minimal number of latent variables are

Table 1. The results of MLR analysis with different types of descriptors.

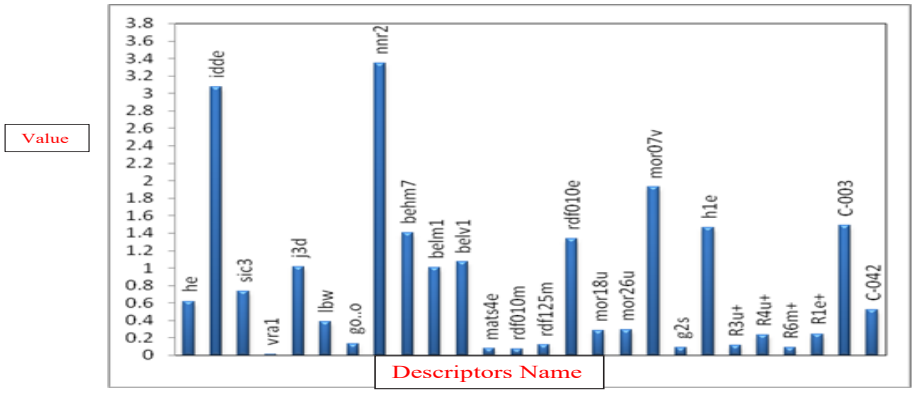
| No. | Descriptor Source | MLR Equations | N | $R^2$ | SE | $Q^2$ | F |
|---|---|---|---|---|---|---|---|
| $E_1$ | Chemical | Y=5.706(±0.267)−0.118(±0.018)HE−0.132(±0.052) LOGP | 140 | 0.31 | 0.27 | 0.28 | 30.67 |
| $E_2$ | Constitutional | Y=−11.686(±2.421)+24.864(±3.740) MV+0.486(±0.088)NN−0.490(±0.179) NCL+0.485(±0.198)NR09 | 140 | 0.52 | 0.31 | 0.48 | 36.9 |
| $E_3$ | Topological | Y=−3.577(±1.066)+3.418(±1.643)SIC3+0.033(±0.005) T(N..N)+0.057(±0.012)PCR−0.018(±0.007) T(O..O)+2.846(±0.770)AAC−0.002(±0.001) VRA1+1.148(±0.341)IDDE−0.026(±0.012)S0K | 140 | 0.65 | 0.29 | 0.56 | 28.43 |
| $E_4$ | Geometrical | Y=12.426(±1.010)+0.032(±0.006) G(N..N)−2.222(±0.345)J3D−0.039(±0.010) G(O..O)+0.181(±0.046)L/BW−0.332(±0.117) SPAN+0.011(±0.005)G(N..O) | 140 | 0.56 | 0.45 | 0.34 | 28.31 |
| $E_5$ | Functional | Y=6.139(±0.169)−0.336(±0.074)NCS+0.892(±0.165) NCN+0.852(±0.124)NNR2+0.850(±0.264) NOHPH−0.315(±0.148)NCRHR+1.100(±0.403) NRSR+0.836(±0.246)nN-NPh−0.660(±0.252) n=CHR−0.361(±0.148)NNRORPH | 140 | 0.66 | 0.41 | 0.54 | 27.91 |
| $E_6$ | Charge | Y=8.818(±0.446)−0.106(±0.020)PCWTE | 140 | 0.17 | 0.47 | 0.15 | 29.0 |
| $E_7$ | Mol-walk | Y=4.440(±0.487)+0.069(±0.028) SRW05+2.087(±0.836)MWC08 | 140 | 0.13 | 0.51 | 0.11 | 9.87 |
| $E_8$ | BCUT | Y=−14.732(±9.786)−34.236(±3.595) BELM1+22.804(±4.381)BELV1+14.516(±2.662) BEHE2−4.548(±1.126)BEHM7+0.223(±0.104) BEHM1 | 140 | 0.5 | 0.34 | 0.44 | 26.8 |
| $E_9$ | Galves | Y=3.434(±0.561)+7.623(±1.989)JGT−12.543(±2.042) GGI10+4.555(±1.380)GGI7−2.437(1.023)GGI6 | 140 | 0.33 | 0.17 | 0.27 | 16.63 |
| $E_{10}$ | 2D | Y=−8.911(±2.627)+12.938(±2.867) ATS3V−5.970(±1.003)MATS4E+29.101(±6.171) ATS3E−20.449(±5.134)ATS7E | 140 | 0.5 | 0.22 | 0.44 | 33.62 |
| $E_{11}$ | RDF | Y=6.071(±0.528)−0.113(±0.017)RD-F070U+0.183(±0.037)RDF030M−1.781(±0.419) RDF020V | 140 | 0.58 | 0.33 | 0.37 | 20.1 |
| $E_{12}$ | 3D | Y=4.989(±0.717)+0.146(±0.053)MO-R03U−0.401(±0.153)MOR09U−2.514(±0.606) MOR26V+1.462(±0.415)MOR27U+0.633(±0.269) MOR18U+0.811(±0.181)MOR07V−0.003(±0.001) MOR01U−0.282(±0.124)MOR05U | 140 | 0.64 | 0.3 | 0.56 | 28.83 |
| $E_{13}$ | WHIM | Y=8.493(±4.450)−115.654(±26.067) G2S+99.551(±27.275)G1P+9.618(±3.572)DV | 140 | 0.17 | 0.38 | 0.14 | 8.99 |
| $E_{14}$ | Getaway | Y=−2.161(±2.188)+19.076(±8.382) R5v++7.339(±1.002)H1e−2.087(±0.419) H3e−1.276(±0.369)H2e−9.360(±2.224) R6m++15.791(±3.636)R3p+−8.084(±2.239) R1e+−12.778(±4.468)R3u++6.237(±2.232) R4e++0.014(±0.007)ITH | 140 | 0.65 | 0.19 | 0.57 | 24.06 |
| $E_{15}$ | Atom-center | Y=7.530(±0.420)−1.860(±0.173)C-003+0.895(±0.137) N-072+0.486(±0.098)C-034−0.059(±0.027) H-047+0.609(±0.165)C-040−0.402(±0.159) C-042+1.270(±0.386)S-107−0.737(±0.313)C -004+0.652(±0.288)C-039 | 140 | 0.71 | 0.31 | 0.63 | 36.36 |
| $E_{16}$ | All molecular descriptor | Y=3.414(±4.378)+9.959(±4.374)MV+0.409(±0.087) NN+0.640(±0.105)NNR2+0.648(±0.218) NOHPH−0.469(±0.136)NCL−5.576(±2.259) BELV1+0.638(±0.210)IDDE−2.489(±0.796) MATS4E−0.004(±0.001)VRA1+0.650(±0.134) SPAN−1.075(±0.285)RDF020V+9.371(±3.563) R3p+−0.388(±0.151)NCRHR | 140 | 0.76 | 0.39 | 0.71 | 30.57 |

Table 2. Correlation coefficient (R2) matrix for the descriptors of azole derivatives used in the MLR equation.

| Descriptors | MV | NN | NNR2 | NOHPH | NCL | BELV1 | IDDE | MATS4E | VRA1 | SPAN | RDF020V | R3p+ | NCRHR |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MV | 1 | .418 | .302 | .182 | .298 | -.104 | .361 | -.064 | -.068 | -.195 | -.444 | .574 | -.231 |
| NN | | 1 | -.180 | .478 | .160 | -.321 | .185 | -.158 | -.093 | .107 | -.118 | .140 | -.237 |
| NNR2 | | | 1 | -.173 | .003 | -.068 | .261 | -.043 | .078 | -.299 | -.248 | .279 | .003 |
| NOHPH | | | | 1 | .077 | .021 | .128 | -.042 | -.143 | .088 | .009 | .055 | .273 |
| NCL | | | | | 1 | -.184 | .159 | .131 | -.069 | -.016 | -.110 | .171 | -.109 |
| BELV1 | | | | | | 1 | -.080 | .080 | .207 | .171 | .148 | -.240 | .362 |
| IDDE | | | | | | | 1 | .257 | .235 | .284 | -.056 | .117 | -.090 |
| MATS4E | | | | | | | | 1 | .279 | .275 | .033 | -.209 | .172 |
| VRA1 | | | | | | | | | 1 | .571 | .063 | -.201 | -.059 |
| SPAN | | | | | | | | | | 1 | .325 | -.409 | -.194 |
| RDF020V | | | | | | | | | | | 1 | -.067 | .148 |
| R3p+ | | | | | | | | | | | | 1 | -.143 |
| NCRHR | | | | | | | | | | | | | 1 |

used for modeling in PLS; this modeling method coincides with noisy data better than MLR. In order to find the more convenient set of descriptors in PLS modeling, genetic algorithm was used as a


A




B

Figure 1. A) PLS regression coefficients for the variables used in GA-PLS model. B) Variable importance in the projection (VIP) for the variables used in GA-PLS model.

feature selection method. To do so, many different GA-PLS runs were conducted using different initial sets of population. The data set (n=200) was divided into two group: calibration set (n=140) and prediction set (n=60). Given 140 calibration samples, the leave-one out cross-validation procedure was used to find the optimum number of latent variables for each PLS model. The most convenient GA-PLS model that resulted in the best fitness contained 26 indices, six of which were those obtained by MLR. The PLS estimate of coefficients for these descriptors are given in Figure 1B. As observed, a combination of chemical, topological, geometrical, functional, BCUT, 2D, 3D, RDF, WHIM, and GETAWAY descriptors have been selected by GA-PLS to account the aromatase inhibitory activity of azole derivatives. The resulted GA-PLS model possessed a high statistical quality $R_2$=0.86 and $Q_2$=0.83. The predictive ability of the model was measured by applying this model to 60 external test set molecules. The squared correlation coefficient for prediction was 0.87, and the standard error of prediction was 0.19.

To measure the significance of the 25 selected PLS descriptors in the aromatase inhibitory activity; variable importance in the projection (VIP) was calculated for each descriptor (39). The VIP analysis of PLS equation is shown in Figure 2. VIP shows that IDDE, NNR2, MOR07, VBEHM7, RDF010E, H1E, and C-003 are the most important indices in the QSAR equation derived by PLS analysis. In addition, J3O, BELM1, and BELV1 have been found to be the moderately influential parameters.

### *3.3. FA-MLR and PCRA*

Table 3 shows the seven factor loadings of the variables (after VARIMAX rotation) for the compounds. Moreover, statistical parameters for testing the prediction ability of the MLR, GA-PLS, PCR, FA-MLR models are provided in Table 4.

As can be seen in Table 3, about 70% of variances in the original data matrix could be explained by selected seven factors. Based on the procedure explained in the experimental section,

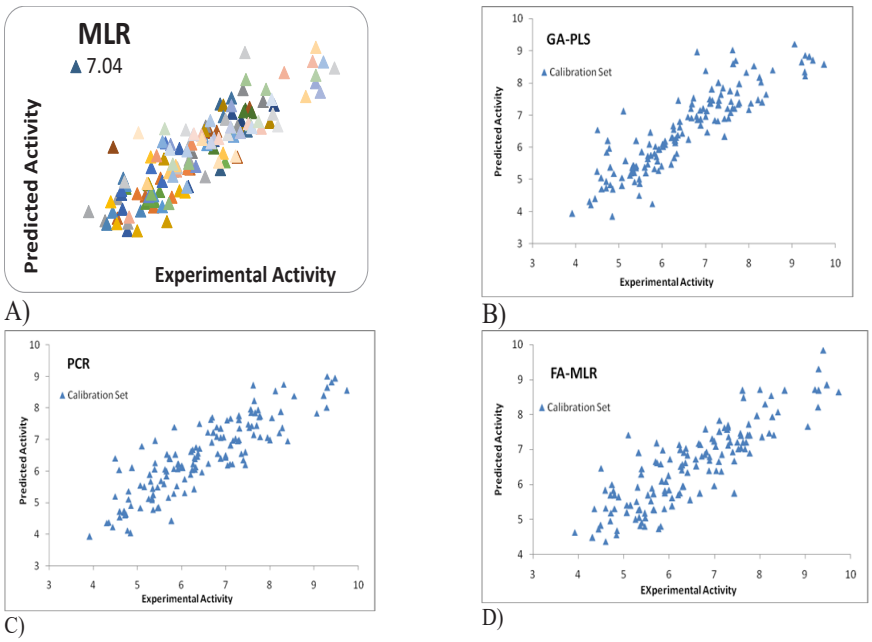



Figure 2. Plots of the cross-validated predicted activity against the experimental activity for the QSAR models obtained by different chemometrics methods

Table 3. Numerical values of factor loading numbers 1-7 for some descriptors after VARIMAX rotation.

| Descriptors | Component | | | | | | | Extraction |
|---|---|---|---|---|---|---|---|---|
| | $F_1$ | $F_2$ | $F_3$ | $F_4$ | $F_5$ | $F_6$ | $F_7$ | |
| HE | .862 | .046 | .027 | .099 | .250 | .038 | .014 | .820 |
| IDDE | -.050 | .003 | .676 | -.166 | -.120 | .193 | -.152 | .561 |
| SIC3 | -.398 | -.395 | .532 | -.317 | .214 | .070 | .142 | .769 |
| VRA1 | .093 | .287 | .029 | .082 | -.016 | .467 | .556 | .625 |
| J3D | .149 | .085 | -.018 | -.049 | .112 | .064 | -.741 | .598 |
| LBW | .043 | .018 | .049 | -.033 | -.872 | .077 | -.036 | .774 |
| GO..O | -.328 | -.207 | -.010 | .038 | .235 | .569 | .301 | .621 |
| NNR2 | .213 | -.034 | .701 | -.098 | .321 | -.142 | .323 | .776 |
| BEHM7 | -.036 | .258 | .243 | .296 | -.183 | .666 | -.021 | .692 |
| BELM1 | .336 | .194 | -.318 | .813 | -.002 | -.017 | -.135 | .931 |
| BELV1 | .046 | -.079 | -.058 | .919 | .032 | .108 | .043 | .870 |
| MATS4E | .250 | -.134 | .149 | -.044 | .020 | .715 | -.095 | .625 |
| RDF010M | .001 | .766 | -.288 | .320 | -.081 | -.024 | -.175 | .810 |
| RDF125M | -.112 | .098 | .009 | -.013 | -.696 | -.091 | .205 | .558 |
| RDF010E | .269 | .738 | -.291 | .334 | -.010 | .047 | -.146 | .837 |
| MOR18U | -.236 | -.274 | .078 | -.625 | -.024 | -.093 | -.330 | .645 |
| MOR26U | .539 | -.133 | -.354 | -.208 | .253 | .198 | -.026 | .581 |
| MOR07V | -.118 | .827 | .266 | -.099 | -.011 | -.004 | .140 | .797 |
| G2S | .233 | .370 | -.225 | -.044 | -.012 | -.169 | .191 | .308 |
| H1E | -.838 | .063 | .064 | -.201 | .188 | -.035 | .097 | .797 |
| R3u+ | .415 | -.577 | -.193 | -.047 | .250 | -.311 | -.022 | .705 |
| R4u+ | .190 | -.482 | .158 | .022 | .455 | -.481 | .066 | .737 |
| R6m+ | -.187 | -.430 | .304 | .069 | -.401 | .086 | -.127 | .502 |
| R1e+ | -.639 | -.322 | -.004 | -.252 | .325 | -.032 | .259 | .749 |
| C-003 | .236 | .125 | -.819 | .091 | .205 | -.207 | .001 | .836 |
| C-042 | -.654 | -.006 | .143 | -.283 | -.302 | .128 | -.100 | .646 |
| %Variance | 13.87 | 12.875 | 10.355 | 10.047 | 9.037 | 7.95 | 5.749 | 69.883 |

the following eight-parametric equation was derived:

Y=16.343(±4.430)−0.811(±0.232)C
-003−0.094(±0.015)HE−2.495(±0.778)MATS4E−4.172
(±2.118)BELM1+0.486(±0.133)NNR2+0.113(±0.036)
L/BW+0.612(±0.231)IDDE− 1.955(±0.817)BEHM
$R_2$=0.77 S.E.=0.47 F=36.10 $Q_2$ =0.73 RMScv=0.34
N=140 (Eq. 1)

Equation 1 could explain 77% of the variance and predict 73% of the variance in (–log$IC_{50}$) data. This equation describes the effect of atom-center (C-003), chemical (HE), 2D (MATS4E), geometrical (L/BW), BCUT (BELM1and BEHM7), functional (NNR2), and topological (IDDE) factors on inhibitory activity.

When factor scores were used as the predictor parameters in a multiple regression equation using forward selection method (PCRA), the following equation was obtained:

Y=6.472(±0.061)+0.688(±0.058) F3−0.573 (±0.060)
F1−0.343 (±0.060) F4−0.313 (±0.058) F6−0.330
(±0.063) F5
$R_2$=0.79 SE=0.25 F=30.1 $Q_2$=0.73 RMScv=0.20
N=140 (Eq. 2)

Equation 2 also shows high equation statistics (79% explained variance and 73% predict

Table 4. Statistical parameters for testing prediction ability of the MLR, GA-PLS, PCR, FA-MLR models.

| Model | $R_2$ | $R_2$LOOCV | RMSEcv | $R_2$p | RMSEp |
|---|---|---|---|---|---|
| **MLR** | 0.76 | 0.71 | 0.29 | 0.78 | 0.15 |
| **GA-PLS** | 0.86 | 0.83 | 0.24 | 0.87 | 0.19 |
| **PCR** | 0.79 | 0.73 | 0.14 | 0.81 | 0.21 |
| **FA-MLR** | 0.77 | 0.73 | 0.32 | 0.78 | 0.15 |

$R_2$: Regression Coefficient for Calibration set

$R_2$LOOCV: Regression Coefficient for Leave One Out Cross Validation

RMSEcv: Root Mean Square Error of cross validation

$R_2$p: Regression Coefficient for prediction set

RMSEp: Root Mean Square Error of prediction set.

variance in (–log$IC_{50}$) data). Since factor scores are used instead of the selected descriptors, and any factor-score contains information from different descriptors, loss of information is thus avoided and the quality of PCRA equation is better than those derived from FA-MLR.

As observed in Table 3, for each factor, the loading values for some descriptors are much higher than those of the others. These high values for some factors indicate that these factors contain higher information about the related descriptor. It should be noted that all factors have information from all descriptors, but the contribution of descriptor in different factors are not equal. For example, factors 1 and 2 have higher loadings for chemical, GETAWAY, WHIM, atom-center, RDF, and 3D indices, whereas information about functional, atom-center, topological, BCUT, and 3D are highly incorporated in factors 3 and 4. Factors 5 and 6 have higher loadings for geometrical, RDF, BCUT, and 2D descriptors and geometrical indices are highly correlated for factor 7. Therefore, the significance of the original variables for modeling the activity can be obtained using the factor scores used by equation $E_2$.

A comparison between the results obtained by GA-PLS and the other employed regression methods indicates the higher accuracy of this method in describing inhibitory activity of the studied compounds (Table 4). Difference in accuracy of the different regression methods used in this study is visualized in Figure 3 by plotting the predicted activity (by cross-validation) against the experimental values. Obviously, all linear models represented scattering of the data around a straight line with slope and intercept close to one and zero, respectively. As it is observed, the plot of data resulted by GA-PLS represents the lowest scattering and those obtained by MLR, FA-MLR, and PCR analysis have lower accuracy.

Table 5. Leverage (h) of the external test set molecules for different models. The last row (h*) is the warning leverage.

| Compound | MLR | GA-PLS | PCR | FA-MLR |
|---|---|---|---|---|
| **2** | 0.14288 | 0.16818 | 0.01900 | 0.05935 |
| **3** | 0.15846 | 0.18979 | 0.01336 | 0.06081 |
| **6** | 0.14079 | 0.44248 | 0.02050 | 0.10788 |
| **8** | 0.05242 | 0.14633 | 0.02284 | 0.02292 |
| **11** | 0.04457 | 0.18829 | 0.01956 | 0.02504 |
| **12** | 0.04496 | 0.18252 | 0.03548 | 0.02472 |
| **15** | 0.08365 | 0.11893 | 0.01119 | 0.05457 |
| **17** | 0.05662 | 0.24540 | 0.04260 | 0.04843 |
| **23** | 0.06222 | 0.10563 | 0.03367 | 0.05171 |

**Continued Table 5.**

| | | | | |
|---|---|---|---|---|
| **24** | 0.05444 | 0.08380 | 0.03471 | 0.04782 |
| **30** | 0.08242 | 0.12324 | 0.02880 | 0.06385 |
| **33** | 0.09269 | 0.13906 | 0.02990 | 0.06190 |
| **36** | 0.15986 | 0.14455 | 0.01827 | 0.02760 |
| **38** | 0.14282 | 0.17320 | 0.02872 | 0.03686 |
| **42** | 0.08951 | 0.31819 | 0.02959 | 0.07830 |
| **49** | 0.08275 | 0.13842 | 0.03434 | 0.05758 |
| **55** | 0.07152 | 0.23245 | 0.01505 | 0.05855 |
| **60** | 0.07627 | 0.16194 | 0.03667 | 0.05823 |
| **62** | 0.08087 | 0.11927 | 0.01121 | 0.04788 |
| **64** | 0.06017 | 0.12505 | 0.01246 | 0.05382 |
| **68** | 0.08534 | 0.13683 | 0.02565 | 0.05306 |
| **75** | 0.04619 | 0.22383 | 0.01451 | 0.05328 |
| **78** | 0.04814 | 0.11056 | 0.01386 | 0.04680 |
| **79** | 0.07264 | 0.14557 | 0.01775 | 0.04778 |
| **80** | 0.07120 | 0.14603 | 0.01072 | 0.06557 |
| **86** | 0.06729 | 0.25212 | 0.02520 | 0.05582 |
| **88** | 0.08297 | 0.18762 | 0.01224 | 0.06388 |
| **90** | 0.03553 | 0.11136 | 0.01334 | 0.04988 |
| **93** | 0.03517 | 0.12404 | 0.00693 | 0.05060 |
| **94** | 0.06407 | 0.07648 | 0.01151 | 0.05180 |
| **100** | 0.07189 | 0.21451 | 0.01538 | 0.07079 |
| **101** | 0.08671 | 0.28550 | 0.03191 | 0.06863 |
| **106** | 0.02817 | 0.11374 | 0.01337 | 0.02205 |
| **107** | 0.04614 | 0.14807 | 0.00290 | 0.04854 |
| **120** | 0.03101 | 0.25959 | 0.03490 | 0.09864 |
| **122** | 0.05894 | 0.15148 | 0.04893 | 0.07139 |
| **123** | 0.03172 | 0.19658 | 0.03765 | 0.07707 |
| **124** | 0.02772 | 0.15548 | 0.01346 | 0.05116 |
| **126** | 0.03118 | 0.18710 | 0.02195 | 0.08277 |
| **127** | 0.04175 | 0.23179 | 0.01327 | 0.05130 |
| **129** | 0.15895 | 0.20530 | 0.05681 | 0.10465 |
| **136** | 0.09792 | 0.17117 | 0.06216 | 0.08211 |
| **138** | 0.06488 | 0.24813 | 0.04207 | 0.10430 |
| **139** | 0.04863 | 0.17015 | 0.03389 | 0.07303 |
| **141** | 0.07266 | 0.14159 | 0.03156 | 0.03451 |
| **143** | 0.05513 | 0.13488 | 0.03617 | 0.05047 |
| **147** | 0.15681 | 0.29189 | 0.01387 | 0.06859 |
| **150** | 0.08071 | 0.09786 | 0.03842 | 0.05394 |
| **151** | 0.11455 | 0.11796 | 0.07559 | 0.07918 |
| **152** | 0.11423 | 0.39927 | 0.03597 | 0.09087 |
| **160** | 0.08094 | 0.09714 | 0.03402 | 0.05666 |
| **164** | 0.10503 | 0.17022 | 0.03800 | 0.09241 |

**Continued Table 5.**

| | | | | |
|---|---|---|---|---|
| **168** | 0.06442 | 0.35440 | 0.05685 | 0.07659 |
| **173** | 0.04802 | 0.23025 | 0.02384 | 0.05663 |
| **176** | 0.07474 | 0.18685 | 0.05295 | 0.01904 |
| **177** | 0.08959 | 0.24565 | 0.05915 | 0.03364 |
| **178** | 0.08159 | 0.25553 | 0.06755 | 0.03956 |
| **179** | 0.07051 | 0.13105 | 0.07381 | 0.03773 |
| **193** | 0.04646 | 0.14804 | 0.01009 | 0.05729 |
| **195** | 0.06454 | 0.36498 | 0.01090 | 0.10982 |
| **h*** | 0.27857 | 0.55714 | 0.10714 | 0.17143 |

### *3.4. Applicability domain of the models*

The calculated leverage values of the test set samples for different MLR, GA-PLS, PCR, and FA-MLR models are listed in Table 5. The warning leverage, as the threshold value for the accepted prediction, is also given in Table 5.

It is important to emphasize that no matter how valid and significant a QSAR may be, it cannot be expected to reliably predict the modeled property for the entire space of chemicals. Therefore, before a QSAR is put into use for screening chemicals, its domain of application must be defined, and predictions for only those chemicals that fall within this domain may be considered reliable. Leverage (32) is one of standard methods for this aim. The numerical value of leverage has certain characteristic: (a) the value is always greater than zero, (b) the lower the value; the higher is the confidence in the prediction. A value of 1 indicates a very poor prediction. A value of 0 indicates perfect prediction and will not be achieved. Another factor for analysis of the results is warning leverage (h*). The warning leverage is, generally, fixed at 3k/n, where n is the number of training compounds, and k is the number of final model parameters. A leverage greater than warning leverage h* means that the predicted response is the result of substantial extrapolation of the model and therefore may not be reliable. As can be seen in Table 5, the leverages of all test samples are lower than h* for all models. This means that all predicted values are acceptable.

### *3.5. Y-randomization test*

The results of this part of study including the R2 and R2LOOCV values after several Y-randomization tests are presented in Table 6.

The robustness of the obtained QSAR models was confirmed by Y-randomization test (40). In this technique, the dependent variable vector is randomly displaced ten times and new

Table 6. R2 and R2LOOCV values after several Y-randomization on all models.

| Iteration | FA-MLR | | PCR | | GA-PLS | | MLR | |
|---|---|---|---|---|---|---|---|---|
| | R2 | R2LOOCV | R2 | R2LOOCV | R2 | R2LOOCV | R2 | R2LOOCV |
| **1** | 0.07 | 0.05 | 0.09 | 0.00 | 0.04 | 0.01 | 0.12 | 0.01 |
| **2** | 0.05 | 0.01 | 0.22 | 0.19 | 0.05 | 0.01 | 0.26 | 0.14 |
| **3** | 0.12 | 0.06 | 0.03 | 0.01 | 0.03 | 0.01 | 0.23 | 0.09 |
| **4** | 0.07 | 0.02 | 0.26 | 0.11 | 0.27 | 0.15 | 0.08 | 0.03 |
| **5** | 0.04 | 0.02 | 0.05 | 0.01 | 0.11 | 0.06 | 0.14 | 0.08 |
| **6** | 0.06 | 0.01 | 0.13 | 0.04 | 0.05 | 0.02 | 0.09 | 0.04 |
| **7** | 0.09 | 0.02 | 0.16 | 0.08 | 0.12 | 0.07 | 0.11 | 0.08 |
| **8** | 0.03 | 0.01 | 0.14 | 0.06 | 0.07 | 0.03 | 0.06 | 0.03 |
| **9** | 0.06 | 0.04 | 0.11 | 0.09 | 0.02 | 0.00 | 0.05 | 0.01 |
| **10** | 0.13 | 0.09 | 0.13 | 0.07 | 0.15 | 0.07 | 0.03 | 0.00 |

QSAR models are developed using the original independent variable matrix. Lower R2 and R2CV-LOO values for the new QSAR models insure the robustness of the obtained QSAR models for the specific modeling method and data (Table 6).

## 4. Conclusion

Quantitative relationships between molecular structure and the inhibitory activity of a series of azole derivative were discovered by a collection of chemometrics methods including MLR, GA-PLS, FA-MLR, and PCRA. In this series, a significant role of topological, functional, 3D, atom-center, and getaway parameters on the inhibitory activity was observed. A comparison between the different employed statistical methods indicated that GA-PLS represented superior results, and it could explain and predict 86% and 83% of variances in the –log $IC_{50}$ data. As observed, the plot of data resulted by GA-PLS represents the lowest scattering, and the impact of topological and functional descriptors were the most.

## Conflict of Interest

None declared.

## 5. References


1. Ortiz AR, Pastor M, Palomer A, Cruciani G, Gago F, Wade RC. Reliability of comparative molecular field analysis models: effects of data scaling and variable selection using a set of human synovial fluid phospholipase A2 inhibitors. *J Med Chem.* 1997;40:1136-48.
2. Ortiz AR, Pisabarro MT, Gago F, Wade RC. Prediction of drug binding affinities by comparative binding energy analysis. *J Med Chem.* 1995;38:2681-91.
3. Bennion C, Connolly S, Gensmantel NP, Hallam C, Jackson CG, Primrose WU, Roberts GC, Robinson DH, Slaich PK. Primrose, et al. Design and synthesis of some substrate analog inhibitors of phospholipase A2 and investigations by NMR and molecular modeling into the binding interactions in the enzyme-inhibitor complex. *J Med Chem.* 1992;35:2939-51.
4. Ortiz AR, Pisabarro MT, Gallego J, Gago F. Implications of a consensus recognition site for phosphatidylcholine separate from the active site in cobra venom phospholipases A2. Biochemistry. 1992;31:2887-96
5. Sessions RB, Dauber-Osguuthorpe P, Campbell MM, Osguthorpe DJ. Modeling of substrate and inhibitor binding to phospholipase A2. *Proteins.* 1992;14:45-64.
6. Noel JP, Bingman CA, Deng TL, Dupureur CM, Hamilton KJ, Jiang RT, et al. Phospholipase A2 engineering. X-ray structural and functional evidence for the interaction of lysine-56 with substrates. *Biochemistry.* 1991;30:11801-11.
7. Chumsri S, Howes T, Bao T, Sabnis G, Brodie A. Aromatase, aromatase inhibitors, and breast cancer. *J Steroid Biochem Mol Biol.* 2011;125:13-22.
8. Hong Y, Chen S. Aromatase inhibitors: structural features and biochemical characterization. *Ann N Y Acad Sci.* 2006;1089:237-51.
9. Séralini G, Moslemi S. Aromatase inhibitors: past, present and future. Mol Cell Endocrinol. 2001;178:117-31.
10. Foye WO, Lemke TL, Williams DA. Foye's principles of medicinal chemistry. Lippincott Williams & Wilkins. 2008;1334-1400.
11. Miller WR1, Bartlett J, Brodie AM, Brueggemeier RW, di Salle E, Lønning PE, et al. Aromatase inhibitors: are there differences between steroidal and nonsteroidal aromatase inhibitors and do they matter? *Oncologist.* 2008;13:829-37.
12. Narashimamurthy J1, Rao AR, Sastry GN. Aromatase inhibitors: a new paradigm in breast cancer treatment. *Curr Med Chem Anticancer Agents.* 2004;4:523-34.
13. Beale JM, Block J, Hill R. Organic medicinal and pharmaceutical chemistry. Lippincott Williams & Wilkins Philadelphia 2010;156-200.
14. Deeb O, Clare BW. QSAR of aromatic substances: protein tyrosine kinase inhibitory activity of flavonoid analogues. *Chem Biol Drug Des*. 2007;70:437-49.
15. Todeschini R. Milano chemometrics and QSPR Group. 2008.
16. Castellano S, Stefancich G, Ragno R, Schewe K, Santoriello M, Caroli A, et al. CYP19 (aromatase): exploring the scaffold flexibility for novel selective inhibitors. *Bioorg Med Chem.* 2008;16:8349-58.
17. Wang R, Shi HF, Zhao JF, He YP, Zhang HB, Liu JP. Design, synthesis and aromatase in-

hibitory activities of novel indole-imidazole derivatives. *Bioorganic Med Chem Lett.* 2013;23:1760-2.
18. Woo LW1, Sutcliffe OB, Bubert C, Grasso A, Chander SK, Purohit A, et al. First dual aromatase-steroid sulfatase inhibitors. *J Med Chem.* 2003;46:3193-6.
19. Gobbi S, Cavalli A, Negri M, Schewe KE, Belluti F, Piazzi L, et al. Imidazolylmethylbenzophenones as highly potent aromatase inhibitors. *J Med Chem*. 2007;50:3420-2.
20. Yahiaoui S, Pouget C, Buxeraud J, Chulia AJ, Fagnère C. Lead optimization of 4-imidazolylflavans: New promising aromatase inhibitors. *Eur J Med Chem.* 2011;46:2541-5.
21. Sonnet P, Dallemagne P, Guillon J, Enguehard C, Stiebing S, Tanguy J, et al. New aromatase inhibitors. Synthesis and biological activity of aryl-substituted pyrrolizine and indolizine derivatives. *Bioorg Med Chem.* 2000;8:945-55.
22. Gobbi S, Zimmer C, Belluti F, Rampa A, Hartmann RW, Recanatini M, et al. Novel highly potent and selective nonsteroidal aromatase inhibitors: synthesis, biological evaluation and structure-activity relationships investigation. *J Med Chem*. 2010;53:5347-51
23. Nagar S, Islam MA, Das S, Mukherjee A, Saha A. Pharmacophore mapping of flavone derivatives for aromatase inhibition. *Mol Divers.* 2008;12:65-76.
24. Saberi MR, Vinh TK, Yee SW, Griffiths BJ, Evans PJ, Simons C. Potent CYP19 (aromatase) 1-[(benzofuran-2-yl)(phenylmethyl) pyridine,-imidazole, and-triazole inhibitors: synthesis and biological evaluation. *J Med Chem.* 2006;49:1016-22.
25. Pouget C, Yahiaoui S, Fagnere C, Habrioux G, Chulia AJ. Synthesis and biological evaluation of 4-imidazolylflavans as nonsteroidal aromatase inhibitors. *Bioorg Chem.* 2004;32:494-503.
26. Lézé MP, Le Borgne M, Pinson P, Palusczak A, Duflos M, Le Baut G, et al. Synthesis and biological evaluation of 5-[(aryl)(1H-imidazol-1-yl) methyl]-1H-indoles: potent and selective aromatase inhibitors. *Bioorg Med Chem Lett.* 2006;16:1134-7.
27. Hackett JC, Kim YW, Su B, Brueggemeier RW. Synthesis and characterization of azole isoflavone inhibitors of aromatase. *Bioorg Med Chem.* 2005;13:4063-70.
28. Yahiaoui S, Pouget C, Fagnere C, Champavier Y, Habrioux G, Chulia AJ. Synthesis and evaluation of 4-triazolylflavans as new aromatase inhibitors. *Bioorg Med Chem Lett.* 2004;14:5215-8.
29. Lézé MP, Palusczak A, Hartmann RW, Le Borgne M. Synthesis of 6-or 4-functionalized indoles via a reductive cyclization approach and evaluation as aromatase inhibitors. *Bioorg Med Chem Lett.* 2008;18:4713-5.
30. Karjalainen A, Kalapudas A, Södervall M, Pelkonen O, Lammintausta R. Synthesis of new potent and selective aromatase inhibitors based on long-chained diarylalkylimidazole and diarylalkyltriazole molecule skeletons. *Eur J Pharm Sci.* 2000;11:109-31.
31. Woo LW1, Bubert C, Sutcliffe OB, Smith A, Chander SK, Mahon MF, et al. Dual aromatase-steroid sulfatase inhibitors. *J Med Chem.* 2007;50:3540-60.
32. Kennard KW, Stone LA. Computer Aided Design of Experiments. *Technometrics.* 1969;11:137-48.
33. Bhattacharya P, Roy K. QSAR of adenosine A3 receptor antagonist 1, 2, 4-triazolo [4, 3-a] quinoxalin-1-one derivatives using chemometric tools. *Bioorg Med Chem Lett.* 2005;15:3737-43.
34. Leardi R. Genetic algorithms in chemometrics and chemistry: a review. *J Chemometrics.* 2001;15:559-69.
35. Hemmateenejad B. Optimal QSAR analysis of the carcinogenic activity of drugs by correlation ranking and genetic algorithm-based PCR. *J Chemometrics.* 2004;18:475-85.
36. Franke R, Gruska A, Waterbeemd H. Chemometrics Methods in molecular design. Methods and Principles in Medicinal Chemistry 1995;2:113-9.
37. H. Kubinyi, 'The quantitative analysis of structure-activity relationships'. w Wolff, M.(red.), Burger's Chemistry and Drug Discovery. John Wiley & Sons Inc., New York 1995.
38. Olah M, Bologa C, Oprea TI. An automated PLS search for biologically relevant QSAR descriptors. *J Comput Aided Mol Des.* 2004;18:437-49.
39. Brereton RG. Applied chemometrics for scientists. John Wiley & Sons. 2007.
40. Baumann K. Cross-validation as the objective function for variable-selection techniques. *Trends Analyt Chem.* 2003;22:395-406.