QSAR Study of Anthranilic Acid Sulfonamides as Inhibitors of Methionine Aminopeptidase-2 using different chemometrics tools

Pooria Zare[1,2];PhD (ID), Maryam Sabet[3];PhD (ID), Razieh Sabet[4*];PhD; (ID)

[1]Department of Pathology, School of Medicine, Kermanshah University of Medical Sciences, Kermanshah, Iran
[2]Department of Master in Public Health, Faculty of Pharmacy, School of Medicine, Shiraz University of Medical Sciences, Shiraz, Iran
[3]Department of Computer Engineering, Shiraz Branch, Islamic Azad University, Shiraz, Iran.
[4]Department of Medicinal Chemistry, School of Pharmacy, Shiraz University of Medical Sciences, Shiraz, Iran

Abstract

Quantitative structure activity relationships (QSAR) studies, as one of the most important areas in chemometrics, play a fundamental role in predicting the biological activity of new compounds and identifying ligand-receptor interactions. Quantitative relationships between molecular structure and methionine aminopeptidase-2 inhibitory activity of a series of anthranilic acid sulfonamides derivatives were discovered by different chemometrics tools including factor analysis based multiple linear regressions (FA-MLR), principale component regression analysis (PCRA) and genetic algorithm-partial least squares GA-PLS. The FA-MLR describes the effect of geometrical and quantum indices on enzyme inhibition activity of the studied molecules. The quality of PCRA equation is better than those derived from FA-MLR. GA-PLS analysis indicated that the topological (IC4 and MPC06), constitutional (nf) and geometrical (G (N..S)) parameters were the most significant parameters on methionine aminopeptidase-2 inhibitory activity. A comparison between the different statistical methods employed revealed that GA-PLS represented superior results and it could explain and predict 85% and 77% of variances in the pIC50 data, respectively.

1. Introduction

Synthesis and evaluation of biological effects of new compounds, usually takes large amounts of time and money. Today the application of computational methods for designing biologically active compounds has opened a new window to modern drug discovery research. Computational methods can accelerate the procedure of discovering new drugs by designing new compounds and predicting potency or activity of them. Quantitative structure activity relationships (QSAR) studies, as one of the most important areas in chemo-

metrics, play a fundamental role in predicting the biological activity of new compounds and identifying ligand-receptor interactions (1-5). QSAR models are mathematical equations that provide us a deeper knowledge about the mechanism of biological activity of compounds by constructing a relationship between chemical structures and biological activities. The most important step in building QSAR models is the appropriate representation of the structural and physicochemical features of chemical entities (6-9). These features called molecular descriptors are the ones with higher impact on the biological activity of interest (10-13). Molecular descriptors have been classified into different categories according to differ-

Corresponding Author: Razieh Sabet, Department of Medicinal Chemistry, School of Pharmacy, Shiraz University of Medical Sciences, Shiraz, Iran
Email: rsabet_r@sums.ac.ir

ent approaches including physiochemical, constitutional, geometrical, topological, and quantum chemical descriptors. Dragon and Gaussian are two well-known computational softwares provide us more than 1000 of these descriptors (14,15). At least one of the several variable selection methods including multiple linear regression (MLR), genetic algorithm (GA), partial least squares (PLS), principale component or factor analysis (PCA/FA) might be used in model building step (11-13).

Post-translational modification step is a required step for proper localization and stability of the protein when the polypeptide chain is synthesized in eukaryotic or prokaryotic cells (16). This kind of modifications in eukaryotes begins with the hydrolytic removal of the N-terminal initiator methionines by two major isoforms of a family of metalloprotease enzymes known as methionine aminopeptidases, MetAP-1 and MetAP-2 (17-21). In prokaryotic cells only one of the two isoforms of MetAP removes the formylmethionine N-terminal (22). In eubacteria, for example, Escherichia coli, Bacillus subtilis, and Salmonella typhimurium MetAP-1 isozyme, while in archaea, for example, Methanobacterium thermoautotrophicum, Sulfolobus solfataricus, and Pyrococcus furiosis, MetAP-2 isozyme catalyze this hydrolytic reaction (23).

Angiogenesis is an essential step in the growth and proliferation of cancer cells. Fumagillin, a fungal metabolite, and its derivatives have been reported to be anti-angiogenic agents. They exert these effects through irreversible inhibition of human MetAP-2 enzyme (24-26). It seems that MetAP-2 inhibition could be an approach to the treatment of cancer (27,28). Reversible and irreversible selective inhibitions of MetAP-2 have been reported. A-357300 has demonstrated selective reversible inhibition of MetAP-2 (29). PPI-2458 and CKD-732 are semisynthetic analogues of fumagillin with inhibitory activity in animal models of angiogenesis and tumor growth by irreversible inhibition of MetAP-2 (30,31). Reversible nonselective inhibition of MetAP-1 and MetAP-2 by LAF-389 has been reported (32). Anthranilic acid sulfonamides belong to another class of reversible selective MetAP-2 inhibitors with characteristics suitable for oral administration in human

(33, 34). Since MetAP-2 is a novel target for cancer therapy, it was of our interest to study the quantitative structure-activity relationships of a series of 46 anthranilic acid sulfonamide derivatives reported in literature as inhibitors of MetAP-2 (34). The structural invariants obtained from whole molecular structures and three different chemometrics methods were used to make connections between structural parameters and MetAp-2 inhibitory activity. These methods included factor analysis–MLR (FA-MLR), principal component regression analysis (PCRA) and partial least squares combined with genetic algorithm for variable selection (GA-PLS).

2. Materials and methods
2.1. Activity data and descriptor generation

The biological data used in this study were methionine aminopeptidase-2 inhibitory activity, (in terms of -log IC_{50}), of a set of forty six anthranilic acid sulfonamides derivatives (34). The structural features and biological activity of these compounds are listed in Table 1. and then used for subsequent QSAR analysis as dependent variable. The two-dimensional structures of molecules were drawn using Hyperchem 7.0 software. The final geometries were obtained with the semi-empirical AM1 method in Hyperchem program. The molecular structures were optimized using Polak-Ribiere algorithm until the root mean square gradient was 0.01 kcal mol^{-1}. Some chemical parameters including molecular volume (V), molecular surface area (SA), hydrophobicity (Log P), hydration energy (HE) and molecular polarizability (MP) were calculated using the Hyperchem Software. The resulted geometry by the Hyperchem software was transferred into Dragon program, which was developed by Milano Chemometrics and QSAR Group (14). The Dragon software calculated different functional groups, topological, geometrical and constitutional descriptors for each molecule. Z-matrices of the structures were provided by the Hyperchem software and transferred to Gaussian 98 program. Complete geometry optimization was performed taking the most extended conformation as starting geometries. Semi-empirical molecular orbital calculation (AM1) of the structures was preformed using the Gaussian 98 program (15).

Table 1. Chemical structures of anthranilic acid sulfonamide analogues used in this study and their experimental and predicted activity for MetAP-2 inhibition.
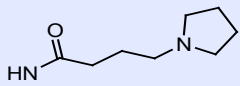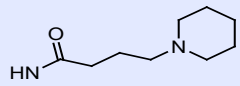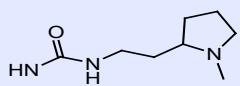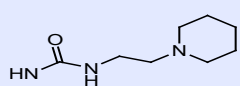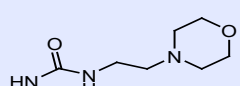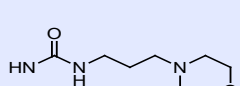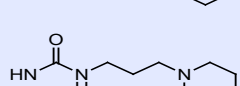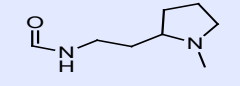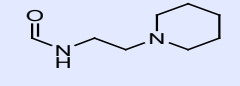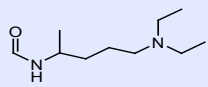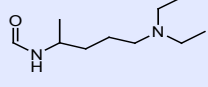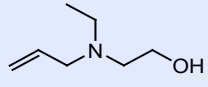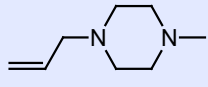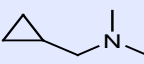

NO.	R1	R2	EXPERIMENTAL PIC50[*] (NM)	PREDICTED PIC50 (NM)		
				PREDICTED PIC50 (NM)	PCR	GA-PLS
1	NHC=OCH2N(CH2CH3)2	H	7.42		7.54	7.37
2	NHC=OCH2CH2N(CH2CH3)2	H	7.23	7.58	7.43	7.38
3[**]	NHC=O(CH2)3N(CH2CH3)2	H	7.23	7.35	7.32	7.33
4		H	7.38	7.23	7.32	7.28
5		H	7.49	7.38	7.42	7.31
6[**]		H	7.14	7.28	7.06	7.48
7[**]		H	7.09	7.25	7.13	7.25
8		H	6.96	7.35	7.06	7.16
9		H	7.26	7.29	7.16	7.18
10[**]		H	7.38	7.35	7.27	7.37
11	NHC=OO(CH2)2N(CH2CH3)2	H	7.29	7.44	7.35	7.25
12[**]	S(CH2)2N(CH2CH3)2	H	7.34	7.50	7.52	7.38
13	S(CH2)3N(CH2CH3)2	H	7.37	7.20	7.39	7.21
14[**]	S(CH2)2N(CH2CH3)2	4-F	7.43	7.41	7.41	7.41
15	S(CH2)2N(CH3)2	4-F	7.52	7.35	7.48	7.53
16	S(CH2)2N(CH3)2	H	7.28	7.28	7.47	7.34
17[**]	SO(CH2)2N(CH2CH3)2	4-F	7.26	7.28	7.20	7.24
18	SO(CH2)2N(CH3)2	4-F	7.21	7.17	7.21	7.31
19	SO(CH2)2N(CH3)2	H	7.27	7.23	7.35	7.28
20	C=ONH(CH2)2N(CH2CH3)2	H	7.51	7.55	7.53	7.41
21[**]		H	7.43	7.68	7.38	7.70
22		H	7.30	7.55	7.40	7.42
23	C=ONH(CH2)3N(CH2CH3)2	H	7.60	7.42	7.44	7.48

Table 1. Continued.

24		4-F	7.36	7.21	7.14	7.29
25		H	7.17	7.19	7.16	7.23
26	CH=CHCH2N(CH2CH3)2	4-F	7.96	7.81	7.81	7.78
27**	CH=CH(CH2)2N(CH2CH3)2	H	7.80	7.86	7.79	7.81
28	CH=CH(CH2)3N(CH2CH3)2	H	7.55	7.51	7.59	7.63
29	CH=CHCH(CH3)N(CH2CH3)2	H	7.72	7.80	7.78	7.69
30	CH=CHCH(CH3)N(CH2CH3)2	4-F	7.80	7.84	7.72	7.86
31**		H	7.82	7.83	7.70	7.88
32		H	7.74	7.66	7.89	7.70
33	CH=CHCH2N(CH3)2	H	7.77	7.76	7.88	7.81
34**	CH=CHCH2N(CH3)2	4-F	7.72	7.76	7.77	7.93
35	CH=CHCH2N(CH3)CH2CH3	H	8.05	7.79	7.85	7.95
36	CH=CHCH2N(CH3)CH2CH2OH	H	7.82	7.82	7.75	8.00
37	CH=CHCH2N(CH3)CH(CH3)2	H	7.92	7.80	7.85	7.94
38	CH=CHCH2N(CH2CH3)2	H	7.57	7.80	7.76	7.75
39**	CH=CH(CH2)2N(CH2CH3)2	H	7.60	7.72	7.74	7.67
40**	CH=CH(CH2)3N(CH2CH3)2	H	7.70	7.51	7.64	7.63
41	(CH2)3N(CH2CH3)2	H	8.05	7.99	7.97	7.92
42	(CH2)3N(CH2CH3)2	4-F	7.77	7.81	7.72	7.87
43**	(CH2)4N(CH2CH3)2	H	7.72	7.79	7.76	7.83
44	(CH2)5N(CH2CH3)2	H	7.35	7.46	7.60	7.51
45	(CH2)3N(CH3)2	H	8.00	7.77	7.93	7.90
46		4-F	7.26	7.50	7.31	7.27

* pIC_{50} = -log (IC_{50})   **Compounds used as prediction set

The Gaussian program calculated different quantum chemical descriptors including, dipole moment (DM), local charges, and HOMO and LOMO energies. Hardness (η), softness (S), electronegativity (χ) and electrophilicity (ω) were calculated according to the method proposed by Thanikaivelan et al (35). The calculated descriptors from whole molecular structures are briefly described in Table 2.

2.2. Data Pretreatment and model building

Methionine aminopeptidase-2 inhibitory activity was used as dependent variable. The cal-culated descriptors (independent variables) were collected in a data matrix whose number of rows and columns were the number of molecules and descriptors, respectively. MLR with factor analysis as the data pre-processing step for variable selection (FA-MLR), principal component regression analysis (PCRA) and Genetic algorithm-partial least squares (GA-PLS) methods were used to derive the QSAR equations.

2.3. Software

A Pentium IV personal computer (CPU at 3.06 GHz) with windows XP operating sys-

Table 2. Brief description of some descriptors used in this study.

Descriptor type	Molecular Description
Constitutional	Molecular weight, no. of atoms, no. of non-H atoms, no. of bonds, no. of heteroatoms, no. of multiple bonds (nBM), no. of aromatic bonds, no. of functional groups (hydroxyl, amine, aldehyde, carbonyl, nitro, nitroso, etc.), no. of rings, no. of circuits, no of H-bond donors, no of H-bond acceptors, no. of Nitrogen atoms (nN), chemical composition, sum of Kier-Hall electrotopological states (Ss), mean atomic polarizability (Mp), number of rotable bonds (RBN), mean atomic Sanderson electronegativity (Me), etc.
Topological	Molecular size index, molecular connectivity indices (X1A, X4A, X2v, X1Av, X2Av, X3Av, X4Av), information content index (IC), Kier Shape indices, total walk count, path/walk-Randic shape indices (PW3, PW4, Zagreb indices, Schultz indices, Balaban J index (such as MSD) Wiener indices, topological charge indices, Sum of topological distances between F..F (T(F..F)), Ratio of multiple path count to path counts (PCR), Mean information content vertex degree magnitude (IVDM), Eigenvalue sum of Z weighted distance matrix (SEigZ), reciprocal hyper-detour index (Rww), Eigenvalue coefficient sum from adjacency matrix (VEA1), radial centric information index, 2D petijean shape index (PJI2), etc.
Geometrical	3D petijean shape index (PJI3), Gravitational index, Balaban index, Wiener index, etc.
Quantum	Highest occupied Molecular Orbital Energy (HOMO), Lowest Unoccupied Molecular Orbital Energy (LUMO), Most positive charge (MPC), Least negative charge (LNC), Sum of squares of charges (SSC), Sum of square of positive charges (SSPC), Sum of square of negative charges (SSNC), Sum of positive charges (SUMPC), Sum of negative charges (SUMNC), Sum of absolute of charges (SAC), Total dipole moment (DMt), Molecular dipole moment at X-direction (DMX), Molecular dipole moment at Y-direction (DMY), Molecular dipole moment at Z-direction (DMZ).
Functional group	Number of total tertiary carbons (nCt), Number of H-bond acceptor atoms (nHAcc), number of total hydroxyl groups (nOH), number of unsubstituted aromatic C(nCaH), number of ethers (aromatic) (nRORPh), etc.
Chemical	LogP (Octanol-water partition coefficient), Hydration Energy (HE), Polarizability (Pol), Molar refractivity (MR), Molecular volume (V), Molecular surface area (SA).

tem was used. Geometry optimization was performed by Hyperchem (version 7.0 Hypercube, Inc.) Dragon software was used for calculation of constitutional, topological, geometrical, functional group descriptors. Gaussian software was used for calculation of quantum descriptors. SPSS software (version 11.50, SPSS, Inc.) was used for PCR and FA-MLR analysis. GA-PLS regression and other calculation were performed in the MATLAB (version 7.1, mathworks, Inc.) environment.

3. Result and discussion
3.1. FA-MLR and PCRA
FA-MLR was 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 vari-

Table 3. Numerical values of factor loading numbers 1–4 for descriptors after VARIMAX rotation.

	1	2	3	4	Commonality
nF	0.133	0.882	0.133	0.054	0.800
PW2	0.026	0.800	0.234	0.070	0.700
SIC2	0.596	0.341	-0.365	0.184	0.639
IC4	0.074	-0.479	0.605	0.284	0.681
MPC06	0.298	0.477	0.727	0.136	0.864
PJI3	-0.032	-0.179	-0.758	0.371	0.745
G(N..S)	0.882	-0.076	0.129	-0.232	0.853
DMx	-0.055	-0.298	0.040	-0.678	0.553
DMz	-0.031	-0.133	-0.025	0.850	0.741
LUMO	-0.671	-0.270	-0.016	-0.323	0.628
PIC50	-0.790	0.055	-0.470	0.056	0.850
%variance	21.139	20.097	17.249	14.742	73.226

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

Model	q^2 [a]	$RMSE_{CV}$ [b]	r^2P [c]	$RMSE_P$ [d]
FA-MLR	0.64	0.17	0.78	0.19
PCR	0.82	0.11	0.82	0.12
GA-PLS	0.77	0.13	0.85	0.14

[a] q^2= Cross validation correlation coefficient;   [b] $RMSE_{CV}$= Root mean square error of cross validation

[c] r^2p= Regression coefficient for prediction set;   [a] $RMSE_P$= Root mean square error of prediction set

ables and to avoid collinearities among them [36]. Principle component regression analysis, PCRA, was 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. PCRA was used for dividing the data set into calibration and prediction set. In this data set there isn't outlier data.

Table 3 shows the four factor loadings of the variables (after VARIMAX rotation) for the compounds tested against Methionine Aminopeptidase-2. As it is observed, about 74% of variances in the original data matrix could be explained by the selected four factors.

Based on the procedure explained in the experimental section, the following three-parametric equation was derived (Equation 1).

pIC_{50}=6.590 (±0.353)-0.054 (±0.007) G (*N..S*) +1.742 (±0.391) PJI3 -0.050 (±0.021) *DMz*

$$r^2 = 0.72\, S.E = 0.21\, F = 29.80\, q^2 = 0.64\, RMScv = 0.17\, N = 38 \quad \text{(Eq. 1)}$$

Equation 1 could explain about 72% of the variance and predict 64% of the variance in pIC50 data. This equation describes the effect of geometrical (G (N..S) and PJI3) and Quantum (DMz) indices on enzyme inhibitory activity of the studied molecules.

When factor scores were used as the predictor parameters in a multiple regression equation using forward selection method (PCRA), the following equation was obtained (Equation 2):

pIC_{50}=7.520(±0.018)-0.215(±0.018)f1-0.144(±0.019)f3

$$r^2 = 0.85\, S.E. = 0.13\, F = 97.82\, q^2 = 0.82\, RMScv = 0.11\, N = 38 \quad \text{(Eq. 2)}$$

Equation 2 could explain and predict 85% and 82% of the variances in pIC50 data, respectively. Since factor scores are used instead of 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 it is observed from Table 3, in the case of each factor, the loading values for some descriptors are much higher than those of the others. These high values for each factor indicate that this factor contains higher information about which descriptors. 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 the geometrical, topological and constitutional indices, whereas information about the topological, geometrical and quantum descriptors are highly incorporated in factor 3 and 4. Therefore, from the factor scores used by equation E2, significance of the original variables for modeling the activity can be obtained. Factor score 1 indicates importance of G (N..S) (Geometrical indice). Factor score 2 indicates importance of nf and PW2 (the constitutional and topological descriptors) and factor score 3 and 4 signify the importance of MPC06, PJI3 and DMz (the topological, geometrical and Quantum descriptors).

The predicted values of the activity for calibration set (by cross-validation) and prediction set for FA-MLR and PCRA are listed in Table 1 and are plotted against the corresponding experimental values in Figure 1. The statistical parameters of prediction set are listed in Table 4. The
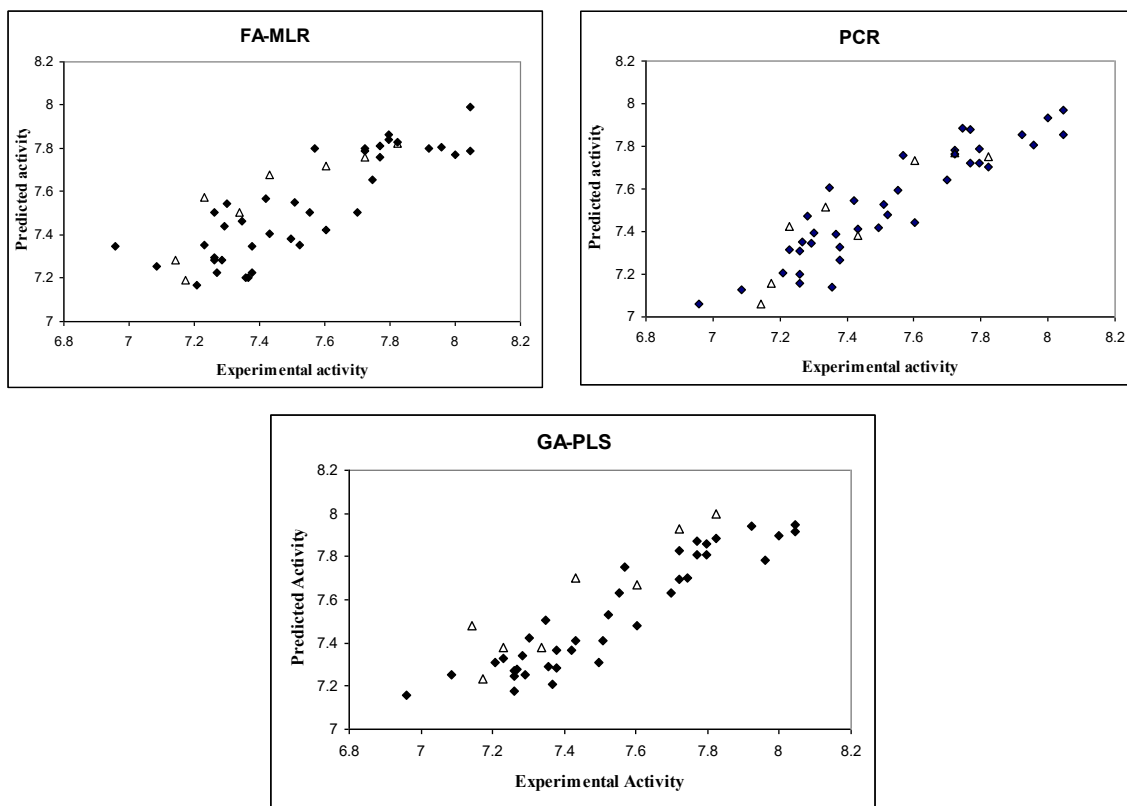
Figure 1. Plots of the cross-validated predicted activity against the experimental activity for the QSAR models obtained by FA-MLR, PCR and GA-PLS methods.

correlation coefficient of prediction for FA-MLR analysis is 0.78, which means that the obtained QSAR model could predict 78% of variances in the MetAP-2 inhibitory activity data. It has a root mean square error of 0.19. The correlation coefficient of prediction for PCRA analysis is 0.82. This means that the derived QSAR model could predict 82% of variances in the inhibitory activity data. The root mean square error of PCRA analysis was 0.12. Whilst the data of this analysis show acceptable prediction, we see that the predicted values of some molecules are near to each other.

3.2. GA-PLS model

In this study, to model the structure-MetAP-2 inhibitory activity relationships better, genetic algorithm-partial least squares (GA-PLS) was employed (38, 39). Application of PLS method thus allows the construction of larger QSAR equations while still avoiding over-fitting and eliminating most variables. This method is normally used

in combination with cross-validation to obtain the optimum number of components (40, 41). The PLS regression method used was the NIPALS-based algorithm existed in the chemometrics toolbox of MATLAB software (version 7.1 Math Work Inc.). In order to obtain the optimum number of factors based on the Haaland and Thomas F-ratio criterion, leave-one-out cross-validation procedure was used (42).

Genetic algorithm is a novel and simple optimization method based on the evolution process of beings in which simplicity and effectiveness have been applied to the various types of optimization problems in many scientific fields. It uses genetic rules such as reproduction, crossover and mutation to build pseudo organisms that are then selected, on the basis of a fitness criterion to survive and pass information on to the next generation (43, 44, 45). Each individual of the population was defined by a chromosome of binary values representing a subset of descriptors. The popula-

Table 5. Leverage (h) of the external test set molecules for different models.

Molecule .no	GA-PLS	PCR	FA-MLR
3	0.14	0.03	0.09
6	0.15	0.06	0.14
7	0.15	0.02	0.06
10	0.10	0.04	0.02
12	0.17	0.05	0.02
14	0.18	0.04	0.04
17	0.21	0.04	0.02
21	0.19	0.05	0.02
27	0.14	0.04	0.05
31	0.14	0.03	0.04
34	0.15	0.02	0.04
39	0.11	0.04	0.05
40	0.15	0.02	0.03
43	0.09	0.08	0.13
h*	0.93	0.18	0.28

The last row (h*) is the warning leverage.

tion size was varied between 50 and 250 for different GA runs. The population of the first generation was selected randomly. The number of genes at each chromosome was equal to the number of descriptors (46). A gene took a value of 1 if its corresponding descriptor was included in the subset; otherwise, it took a value of zero. The number of genes with a value of 1 was kept relatively low to have a small subset of descriptors, that is, the probability of generating 0 for a gene was set greater (at least 70%) than the value of 1 (45). The operators used here were crossover and mutation. The probability of the application of these operators was varied linearly with generation renewal (0-10% for mutation and 60-90% for crossover). For a typical run, the evolution of the generation was stopped when 90% of the generation took the same fitness. A maximum generation number of 500 were used throughout. The fitness function (predictability of the model) was computed by cross-validation procedure based on the sum of squares of errors (SSECV) value. The inverse of SSECV was considered as fitness function (47). The chromosomes with the least numbers of selected descriptors and the highest fitness were marked as informative chromosomes (46).

In PLS analysis, the descriptors data matrix is decomposed to orthogonal matrices with an inner relationship between the dependent and independent variables. The multi-colinearity problem in the descriptors is omitted by PLS analysis because a minimal number of latent variables are used for modeling in PLS (46). Since redundant variables degrade the performance of PLS analysis, similar to other regression methods, a variable selection method must be employed to find the more convenient set of descriptors. Here, GA was used as variable selection method. The data set (n=46) was divided into two group: calibration set (n=38) and prediction set (n=8). Given 38 calibration samples; cross-validation procedure was used to find the optimum number of latent variables for each PLS model. GA produces a population of acceptable models in each run. In this work, many different GA-PLS runs were conducted using different initial set of populations (50-250) and therefore a large number of acceptable models were created.

The most convenient GA-PLS model that resulted in the best fitness contained 10 descriptors including four topological indices (PW2, SIC2, IC4 and MPC06), one constitutional (nf), two geometrical (G (N..S) and PJI3) and three quantum parameters (LUMO, DMz, DMx). The PLS
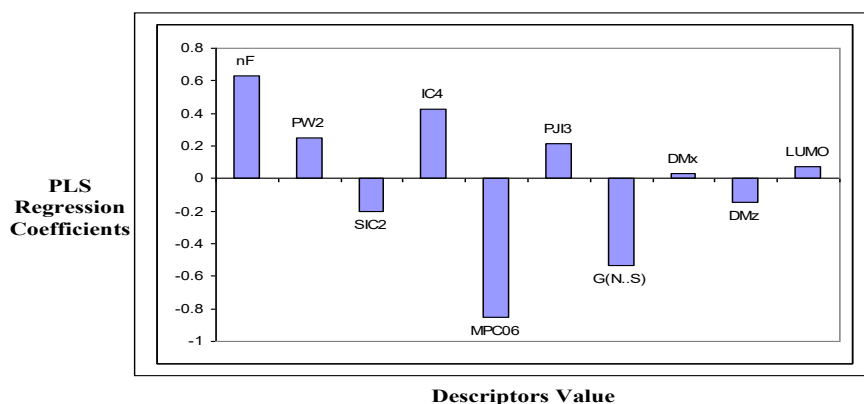
Figure 2. PLS regression coefficients for the variables used in GA-PLS model.

estimate of the regression coefficients are shown in Figure 2. Since these constants were calculated based on the normalized descriptor values, they can be used as a measure of the importance of the corresponding descriptor. As it is observed, the topological (IC4 and MPC06), constitutional (nf) and geometrical (G (N..S)) parameters represent the most significant contribution in the obtained QSAR model followed by the functional geometrical and topological parameters (PJI and SIC2).

The statistical parameters of the resulted PLS-based QSAR model are given in Table 4. The resulted GA-PLS model possessed high statistical quality $R2=0.86$ and $Q2=0.77$. It could explain and predict about 77% of variances in the methionine aminopeptidase-2 inhibitory activity of the studied molecules. The predictive ability of the model was measured by application to 8 external test set molecules. The correlation coefficient of prediction set is 0.85, which means that the resulted QSAR model could predict 85% of variances in the inhibitory activity data and standard error of prediction was 0.13.

The predicted activities are represented in Table 1 and are plotted against the corresponding experimental values in Figure 1. Comparison between the results obtained by GA-PLS and the other employed regression methods indicates higher accuracy of this method in describing inhibitory activity of the anthranilic acid sulfonamides derivative. Difference in accuracy of the different regression methods used in this study is visualized in Figure 1 by plotting the predicted activity (by cross-validation) against the experimental values. As it is observed, the plot of data resulted by

GA-PLS represents the lowest scattering of data around a straight line and that obtained by PCRA analysis is in the second order of accuracy.

3.3 Robustness and applicability domain of the models

Leverage method was used to Robustness and applicability domain of the models, 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. The calculated leverage values of the test set samples for all models and the warning leverage, as the threshold value for accepted prediction, are showed in table 5. As seen, the leverages of all test samples are lower than h* for all models. This means that all predicted values are acceptable (40).

4. Conclusion

Quantitative relationships between molecular structure and methionine aminopeptidase-2 inhibitory activity of a series of anthranilic acid sulfonamides derivatives were discovered by different chemometrics tools including FA-MLR, PCRA and GA-PLS. The FA-MLR describes the effect of geometrical and quantum indices on enzyme inhibitory activity of the studied molecules. The quality of PCRA equation is better than those derived from FA-MLR. Factor score 1 indicates importance of G (N..S) (Geometrical indice). Factor score 2 indicates importance of nf and PW2 (the constitutional and topological descriptors) and factor score 3 and 4 signify the importance of MPC06, PJI3 and DMz (the topological, geometrical and Quantum descriptors).

GA-PLS analysis indicated that the topological (IC4 and MPC06), constitutional (nf) and geometrical (G (N..S)) parameters were the most significant parameters on methionine aminopeptidase-2 inhibitory activity. A comparison between the different statistical methods employed revealed that GA-PLS represented superior results and it could explain and predict 85% and 77% of variances in the pIC50 data, respectively. Compounds 10,15,19,27,46 are the best molecules according to the best QSAR moldels(GA-PLS).

Conflict of Interest

None declared.

References

1. Schmidi H. Multivariate prediction for QSAR. *Chemom Intell Lab Syst.* 1997:3;125-34.

2. Hansch C, Kurup A, Garg R, Gao H. Chem-bioinformatics and QSAR: a review of QSAR lacking positive hydrophobic terms. *Chem Rev.* 2001 Mar;101(3):619-72. doi: 10.1021/cr0000067. PMID: 11712499.

3. Wold S, Trygg J, Berglund A, Antti H. Some recent developments in PLS modeling. *Chemom Intell Lab Syst.* 2001;53:131-50.

4. Hemmateenejad B, Miri R, Akhond M, Shamsipur M. QSAR study of the calcium channel antagonist activity of some recently synthesized dihydropyridine derivatives. An application of genetic algorithm for variable selection in MLR and PLS methods, *Chemom Intell Lab Syst.* 2002;64:1-99.

5. Hemmateenejad B, Miri R, Akhond M, Shamsipur M. Quantitative structure-activity relationship study of recently synthesized 1,4-dihydropyridine calcium channel antagonists. Application of the Hansch analysis method. *Arch Pharm (Weinheim).* 2002 Dec;335(10):472-80. doi: 10.1002/ardp.200290001. PMID: 12506395.

6. Hansch C, Fujita T, Analysis ρ-σ-π. A method for the correlation of biological activity and chemical structure, *J Am Chem Soc.* 1964;86:1616-26.

7. Wan J, Zhang L, Yang G, Zhan CG. Quantitative structure-activity relationship for cyclic imide derivatives of protoporphyrinogen oxidase inhibitors: a study of quantum chemical descriptors from density functional theory. *J Chem Inf Comput Sci.* 2004 Nov-Dec;44(6):2099-105. doi: 10.1021/ci049793p. PMID: 15554680.

8. Hansch C, Hoekman D, Gao H. Comparative QSAR: Toward a Deeper Understanding of Chemicobiological Interactions. *Chem Rev.* 1996 May 9;96(3):1045-1076. doi: 10.1021/cr9400976. PMID: 11848780.

9. R. Todeschini, V. Consonni, Handbook of Molecular Descriptors. Wiley-VCH, Weinheim, 2000.

10. Horvath D, Mao B. Neighborhood behavior. Fuzzy molecular descriptors and their influence on the relationship between structural similarity and property similarity. *QSAR Comb Sci.* 2003;22:498-509.

11. Putta S, Eksterowicz J, Lemmen C, Stanton R. A novel subshape molecular descriptor. *J Chem Inf Comput Sci.* 2003 Sep-Oct;43(5):1623-35. doi: 10.1021/ci0256384. PMID: 14502497.

12. Gupta S, Singh M, Madan AK. Superpendentic index: a novel topological descriptor for predicting biological activity. *J Chem Inf Comput Sci.* 1999 Mar-Apr;39(2):272-7. doi: 10.1021/ci980073q. Erratum in: J Chem Inf Comput Sci 1999 Nov-Dec;39(6):1230. PMID: 10192943.

13. Consonni V, Todeschini R, Pavan M, Gramatica P. Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 2. Application of the novel 3D molecular descriptors to QSAR/QSPR studies. *J Chem Inf Comput Sci.* 2002 May-Jun;42(3):693-705. doi: 10.1021/ci0155053. PMID: 12086531.

14. Dragon, version 2.1, Milano Chemometrics and QSPR Group., Milano, Italy, 2002.

15. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Zakrzewski VG, Montgomery Jr JA, Stratmann RE, Burant JC, Dapprich S. Gaussian 98, revision a. 7, Gaussian. Inc., Pittsburgh, PA. 1998;12.

16. Kendall RL, Yamada R, Bradshaw RA. Cotranslational amino-terminal processing. *Methods Enzymol.* 1990;185:398-407. doi: 10.1016/0076-6879(90)85035-m. PMID: 2143260.

17. Bradshaw RA, Brickey WW, Walker

KW. N-terminal processing: the methionine aminopeptidase and N alpha-acetyl transferase families. *Trends Biochem Sci*. 1998 Jul;23(7):263-7. doi: 10.1016/s0968-0004(98)01227-4. PMID: 9697417.

18. Lowther WT, Matthews BW. Metalloaminopeptidases: common functional themes in disparate structural surroundings. *Chem Rev*. 2002 Dec;102(12):4581-608. doi: 10.1021/cr0101757. PMID: 12475202.

19. Tso JY, Hermodson MA, Zalkin H. Glutamine phosphoribosylpyrophosphate amidotransferase from cloned Escherichia coli purF. NH2-terminal amino acid sequence, identification of the glutamine site, and trace metal analysis. *J Biol Chem*. 1982 Apr 10;257(7):3532-6. PMID: 7037784.

20. Boissel JP, Kasper TJ, Shah SC, Malone JI, Bunn HF. Amino-terminal processing of proteins: hemoglobin South Florida, a variant with retention of initiator methionine and N alpha-acetylation. *Proc Natl Acad Sci U S A*. 1985 Dec;82(24):8448-52. doi: 10.1073/pnas.82.24.8448. PMID: 3866233; PMCID: PMC390933.

21. Chen S, Vetro JA, Chang YH. The specificity in vivo of two distinct methionine aminopeptidases in Saccharomyces cerevisiae. *Arch Biochem Biophys*. 2002 Feb 1;398(1):87-93. doi: 10.1006/abbi.2001.2675. Erratum in: Arch Biochem Biophys. 2003 Sep 1;417(1):128. PMID: 11811952.

22. Frottin F, Martinez A, Peynot P, Mitra S, Holz RC, Giglione C, Meinnel T. The proteomics of N-terminal methionine cleavage. *Mol Cell Proteomics*. 2006 Dec;5(12):2336-49. doi: 10.1074/mcp.M600225-MCP200. Epub 2006 Sep 8. PMID: 16963780.

23. Emami L, Sabet R, Khabnadideh S, Faghih Z, Thayori P. Quinazoline analogues as cytotoxic agents; QSAR, docking, and in silico studies. *Res Pharm Sci*. 2021 Aug 19;16(5):528-546. doi: 10.4103/1735-5362.323919. PMID: 34522200; PMCID: PMC8407157..

24. Griffith EC, Su Z, Turk BE, Chen S, Chang YH, Wu Z, Biemann K, Liu JO. Methionine aminopeptidase (type 2) is the common target for angiogenesis inhibitors AGM-1470 and ovalicin. *Chem Biol*. 1997 Jun;4(6):461-71. doi: 10.1016/s1074-5521(97)90198-8. PMID: 9224570.

25. Sin N, Meng L, Wang MQ, Wen JJ, Bornmann WG, Crews CM. The anti-angiogenic agent fumagillin covalently binds and inhibits the methionine aminopeptidase, MetAP-2. *Proc Natl Acad Sci U S A*. 1997 Jun 10;94(12):6099-103. doi: 10.1073/pnas.94.12.6099. PMID: 9177176; PMCID: PMC21008.

26. Ingber D, Fujita T, Kishimoto S, Sudo K, Kanamaru T, Brem H, Folkman J. Synthetic analogues of fumagillin that inhibit angiogenesis and suppress tumour growth. *Nature*. 1990 Dec 6;348(6301):555-7. doi: 10.1038/348555a0. PMID: 1701033.

27. Folkman J. What is the evidence that tumors are angiogenesis dependent? *J Natl Cancer Inst*. 1990 Jan 3;82(1):4-6. doi: 10.1093/jnci/82.1.4. PMID: 1688381.

28. Folkman J, Watson K, Ingber D, Hanahan D. Induction of angiogenesis during the transition from hyperplasia to neoplasia. *Nature*. 1989 May 4;339(6219):58-61. doi: 10.1038/339058a0. PMID: 2469964.

29. Wang J, Sheppard GS, Lou P, Kawai M, BaMaung N, Erickson SA, et al. Tumor suppression by a rationally designed reversible inhibitor of methionine aminopeptidase-2. *Cancer Res*. 2003 Nov 15;63(22):7861-9. PMID: 14633714.

30. Bernier SG, Lazarus DD, Clark E, Doyle B, Labenski MT, Thompson CD, et al. A methionine aminopeptidase-2 inhibitor, PPI-2458, for the treatment of rheumatoid arthritis. *Proc Natl Acad Sci U S A*. 2004 Jul 20;101(29):10768-73. doi: 10.1073/pnas.0404105101. Epub 2004 Jul 12. PMID: 15249666; PMCID: PMC490009.

31. Chun E, Han CK, Yoon JH, Sim TB, Kim YK, Lee KY. Novel inhibitors targeted to methionine aminopeptidase 2 (MetAP2) strongly inhibit the growth of cancers in xenografted nude model. *Int J Cancer*. 2005 Mar 10;114(1):124-30. doi: 10.1002/ijc.20687. PMID: 15523682.

32. Towbin H, Bair KW, DeCaprio JA, Eck MJ, Kim S, Kinder FR, et al. Proteomics-based target identification: bengamides as a new class of methionine aminopeptidase inhibitors. *J Biol Chem*. 2003 Dec 26;278(52):52964-71. doi: 10.1074/jbc.M309039200. Epub 2003 Oct 8. PMID: 14534293.

33. Kawai M, BaMaung NY, Fidanze SD, Erickson SA, Tedrow JS, Sanders WJ, et al. Development of sulfonamide compounds as potent methionine aminopeptidase type II inhibitors with antiproliferative properties. *Bioorg Med Chem*

Lett. 2006 Jul 1;16(13):3574-7. doi: 10.1016/j.bmcl.2006.03.085. Epub 2006 May 2. PMID: 16632353.

34. Sheppard GS, Wang J, Kawai M, Fidanze SD, BaMaung NY, Erickson SA, et al. Discovery and optimization of anthranilic acid sulfonamides as inhibitors of methionine aminopeptidase-2: a structural basis for the reduction of albumin binding. *J Med Chem*. 2006 Jun 29;49(13):3832-49. doi: 10.1021/jm0601001. PMID: 16789740.

35. Thanikaivelan P, Subramanian V, Rao JR, Nair BU. Application of quantum chemical descriptor in quantitative structure activity and structure property relationship. *Chem Phys Lett*. 2000;323:59-70.

36. Franke R, Gruska A. Chemometrics Methods in molecular design, in: H. van Waterbeemd, (Ed.), Methods and Principles in Medicinal Chemistry, VCH, Weinheim, 1995, Vol. 2, pp. 113–119.

37. Kubinyi H. The quantitative analysis of structure-activity relationships, in: M.E. Wolff, (Ed.), Burger's Medicinal Chemistry and Drug Discovery, 5th Ed.; Wiley, New York, 1995, Vol. 1, pp. 506-509.

38. Leardi R. Application of Genetic Algorithm-PLS for Feature Selection in Spectral Data Sets, J. Chemomtr. 14 (2000) 643-655.

39. Leardi R, Gonzalez AL. Genetic Algorithm Applied to Feature Selection in PLS Regression: How and When to Use Them. *Chemom Intell Lab Syst.* 1998;41:195-207.

40. Fassihi A, Sabet R. QSAR study of p56(lck) protein tyrosine kinase inhibitory activity of flavonoid derivatives using MLR and GA-PLS. *Int J Mol Sci.* 2008 Sep;9(9):1876-1892. doi: 10.3390/ijms9091876. Epub 2008 Sep 22. PMID: 19325836; PMCID: PMC2635749.

41. Leardi R. Genetic Algorithms in Chemometrics and Chemistry: A Review, J. Chemometrics 15 (2001) 559-569.

42. Hemmateenejad B. Optimal QSAR Analysis of the Carcinogenic Activity of Drugs by Correlation Ranking and Genetic Algorithm-Based. *J Chemometrics*. 2004;18:475-485.

43. Cho SJ, Hermsmeier MA. Genetic Algorithm guided Selection: variable selection and subset selection. *J Chem Inf Comput Sci.* 2002 Jul-Aug;42(4):927-36. doi: 10.1021/ci010247v. PMID: 12132894.

44. Ahmad S, Gromiha MM. Design and training of a neural network for predicting the solvent accessibility of proteins. *J Comput Chem*. 2003 Aug;24(11):1313-20. doi: 10.1002/jcc.10298. PMID: 12827672.

45. Faghih Z, Emami L, Zomoridian K, Sabet R, Bargebid R, Mansourian A, et al Aryloxy Alkyl Theophylline Derivatives as Antifungal Agents: Design, Synthesis, Biological Evaluation and Computational Studies. *ChemistrySelect.* 2022;7.

46. Deeb O, Hemmateenejad B, Jaber A, Garduno-Juarez R, Miri R. Effects of the Electronic and Physicochemical Parameters on the Carcinogenesis Activity of Some Sulfa Drug Using QSAR Analysis Based on Genetic-MLR & Genetic-PLS. *Chemosphere*. 2007;67:2122-2130.

47. Salehi F, Emami L, Rezaie Z, Khabnadideh S, Tajik B, Sabet R. Fluconazole-Like Compounds as Potential Antifungal Agents: QSAR, Molecular Docking, and Molecular Dynamics Simulation. *J Chem.* 2022 Mar;2022:1-16