

Article

QSAR Study of p56^{lck} Protein Tyrosine Kinase Inhibitory Activity of Flavonoid Derivatives Using MLR and GA-PLS

Afshin Fassihi* and Razieh Sabet

Department of Medicinal Chemistry, Faculty of Pharmacy, Isfahan University of Medical Sciences and Health Services, 81746-73461, Isfahan, Iran. E-Mail: Sabet@pharm.mui.ac.ir

* Author to whom correspondence should be addressed; E-Mail: fassihi@pharm.mui.ac.ir;
Tel. +98-0311-7922562; Fax: +98-0311-6680011

Received: 9 August 2008; in revised form: 2 September 2008 / Accepted: 13 September 2008 /
Published: 22 September 2008

Abstract: Quantitative relationships between molecular structure and p56^{lck} protein tyrosine kinase inhibitory activity of 50 flavonoid derivatives are discovered by MLR and GA-PLS methods. Different QSAR models revealed that substituent electronic descriptors (SED) parameters have significant impact on protein tyrosine kinase inhibitory activity of the compounds. Between the two statistical methods employed, GA-PLS gave superior results. The resultant GA-PLS model had a high statistical quality ($R^2 = 0.74$ and $Q^2 = 0.61$) for predicting the activity of the inhibitors. The models proposed in the present work are more useful in describing QSAR of flavonoid derivatives as p56^{lck} protein tyrosine kinase inhibitors than those provided previously.

Keywords: Protein tyrosine kinase; Flavonoid; QSAR; Chemometrics, SED analysis.

1. Introduction

The quantitative structure-activity relationship (QSAR) research field provides medicinal chemists with the ability to predict drug activity by mathematical equations which construct a relationship between the chemical structure and the biological activity [1, 2]. These mathematical equations are in the form of $y = Xb + e$ that describe a set of predictor variables (X) with a predicted variable (y) by

means of a regression vector (b) [3]. After the earlier QSAR studies by Hansch, who showed a correlation between biological activity and octanol-water partition coefficient [2], it is now assumed that the sum of substituent effects on the steric, electronic and hydrophobic interaction of compounds with their receptor determines their biological activity [4-6]. The first step in constructing the QSAR models is finding one or more molecular descriptors that represent variation in the structural property of the molecules by a number [7]. Nowadays, a wide range of descriptors are being used in QSAR studies which can be classified into different categories according to the Karelson approach including; constitutional, geometrical, topological, quantum, chemical and so on [8]. There are different variable selection methods available including; multiple linear regression (MLR), genetic algorithm (GA), principal component or factor analysis (PCA/FA) and so on. The mathematical relationships between molecular descriptors and activity are used to find the parameters affecting the biological activity and/or estimate the property of other molecules.

It is now well established that protein tyrosine kinases (PTKs) provide a central switching mechanism in cellular signal transduction pathways by catalyzing the transfer of the γ -phosphate of either ATP or GTP to specific tyrosine residues in certain protein substrates [9, 10]. This regulatory control plays a crucial role in signal transduction pathways that regulate several cellular functions under both normal and deregulated conditions [11-14]. PTKs are the intracellular effectors for many growth hormone receptors. After the discovery of activated PTKs as the product of dominant viral-transforming genes (oncogenes) providing the early hypothesis for the connection between protein tyrosine phosphorylation and cell transformation, enough evidence are now available to suggest that inappropriate or elevated expression of PTKs contribute to the transformed state of cells in many human malignancies [15-19]. P56^{lck} is a lymphoid-specific protein tyrosine kinase that is principally expressed in T lymphocytes [20]. Association of p56lck with the cytoplasmic tail of various cell surface receptors, as well as associations of p56lck with intracellular targets of phosphorylation, suggests that this tyrosine kinase plays a central role in coordinating early signal transduction events [21]. Based on this knowledge it is clear that, substances which can modulate the activity of PTKs might be potentially effective therapeutic agents. The key step in the mechanism of kinase activity of all PTKs is the recognition and binding of a nucleoside triphosphate (usually ATP) and an appropriate tyrosyl-containing substrate to the enzyme. Direct transfer of phosphate between the two molecules is the next step in the PTKs function [22]. A variety of compounds can inhibit the function of PTKs in a manner which is competitive with respect to nucleotide binding. Among such competitive inhibitors are flavonoids, a group of low molecular weight plant natural products that include one of the largest classes of naturally-occurring polyphenolic compounds [23, 24]. This group of plant natural products is largely responsible for the colors of many fruits and flowers, and over 4,000 flavonoid pigments have been characterized and classified according to their chemical structure. Chemically they are C6-C3-C6 compounds in which the two C6 groups are substituted benzene rings, and the C3 group is an aliphatic chain which contains a pyran ring. Flavonoids occur as O- or C-glycosides or in the "free" state as aglycones with hydroxyl or methoxyl groups present on the aglycone. The flavonoids may be divided into seven types: flavones, flavonols, flavonones, chalcones, xanthenes, isoflavones, and biflavones. Flavonoids have been gained wide interest as potential pharmacological agents since some of the best sources of flavonoids are foods: apples, blueberries, bilberries, onions, soy products and tea.

Furthermore numerous medicinal plants contain therapeutic amounts of flavonoids, which are used to treat a wide variety of disorders [25].

Here, we consider the inhibitory activity of flavonoids against protein-tyrosine kinase p56^{lck}. Several QSAR studies were reported on this class of molecules using different descriptors and different methods of modeling. Thakur *et al.* described a QSAR study on p56^{lck} protein tyrosine kinase inhibitor flavonoids using only hydration energy and hydrophobic parameters [26]. Nikolovska-Coleska *et al.* treated a set of 104 derivatives with standard linear regression technique by the use of classical/quantum descriptors [27]. The same dataset was treated by Novic *et al.* with a counter propagation neural network by the use of classical/quantum descriptors [28]. Oblak *et al.* applied a wide variety of descriptors with CODESSA software on the above-mentioned dataset [29]. A quantum chemical/classical QSAR study on a set of 75 flavonoids and closely related compounds tested as p56^{lck} protein tyrosine kinase and AR inhibitors has been carried out by Stefanic *et al.* and the obtained structure-activity relationships of both enzyme systems were compared [30]. A comprehensive *ab initio* study of 3D structures of some flavonoids is reported by Meyer [31]. Deeb *et al.* calculated nodal orientation with program NODANGLE [32].

In the present paper, the QSAR study for a series of 50 flavonoid analogues with the ability to inhibit protein tyrosine kinase has been considered [32]. In a comprehensive study of the PTK system we used a very large descriptor set (more than 600 topological, geometrical, constitutional, functional group, electrostatic, quantum and chemical descriptors) and different analyses: Hansch, Free-Wilson and substituent electronic descriptors (SED), in order to be able to compare the predictive ability of descriptors from different descriptor groups. Multiple linear regression (MLR) and genetic algorithm partial least squares (GA-PLS) methods were applied as methods for modeling.

2. Results and Discussion

The structural features and biological activity of the studied compounds are listed in Table 1. Calculated descriptors for each molecule are summarized in Table 2.

Table 1. Chemical structure of flavonoid derivatives used in this study and their experimental and predicted activity for protein kinase inhibition.

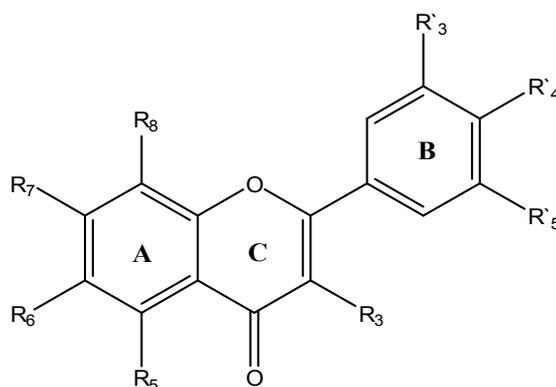


Table 1. Chemical structure of flavonoid derivatives.

Compound	R	Experimental pIC ₅₀ ^a	Predicted pIC ₅₀	REP ^b
1	5,7-OH,4'-NH ₂	5.13	4.7707	-0.0753
2	3,5,7,3',4'-OH	4.88	4.9431	0.0128
3	3,7,3',4'-OH	4.86	4.7707	-0.0187
4	5,7,4'-OH	4.83	4.4356	-0.0889
5	5,4'-OH	4.80	4.2603	-0.1267
6	6,3'-OH	4.80	4.4242	-0.0849
7	6-OH,5,7,4'-NH ₂	4.74	4.1061	-0.1544
8	5,7-OH	4.71	4.0895	-0.1518
9	4'-OH,3',5'-OCH ₃	4.57	4.2687	-0.0706
10	5,7,3',4'-OH	4.46	4.4172	-0.0097
11	7,3'-OH	4.41	4.4358	0.0058
12	6-OH,5,7,3'-NH ₂	4.34	4.3681	0.0064
13	6-OMe,8,3'-NH ₂	4.25	4.1649	-0.0204
14	6-OH,3',4',5'-OCH ₃	4.22	4.3591	0.0319
15	3,5,7,4'-OH,3',5'-OCH ₃	4.16	4.1649	0.0012
16	3,5,7,3',5'-OH	4.00	3.9947	-0.0013
17	6,4'-NH ₂	3.99	3.9613	-0.0072
18	6,8,4'-NH ₂	3.97	3.9764	0.0016
19	6-OH,8,4'-NH ₂	3.93	3.9446	0.0037
20	6,4'-OH	3.93	3.9247	-0.0013
21	7,8,4'-OH,3',5'-OCH ₃	3.92	3.8990	-0.0054
22	8,4'-NH ₂	3.91	3.8994	-0.0027
23	6,4'-OH,3',5'-OCH ₃	3.89	3.9133	0.0060
24	7-OH,4'-NH ₂	3.86	3.8815	0.0056
25	7-OH,6,4'-NH ₂	3.85	3.8296	-0.0053
26	7,4'-OH	3.78	3.8621	0.0213
27	7,8,3'OH	3.75	3.6903	-0.0162
28	6,3'-NH ₂	3.70	4.0228	0.0803
29	4'-NH ₂	3.68	4.1850	0.1207
30	5-OH,6,4'-NH ₂	3.65	3.9325	0.0718
31	3,5,7-OH	3.53	3.9794	0.1129
32	5,4'-OH,7-OCH ₃	3.55	3.7315	0.0487
33	5,3'-OH	3.50	4.1209	0.1507
34	7,8-OH	3.50	3.4873	-0.0036
35	5-OH,8,4'-NH ₂	3.49	3.6705	0.0492
36	7-OH,8,4'-NH ₂	3.48	3.6694	0.0516
37	7-OH	3.47	3.8567	0.1003

Table 1. Cont.

38	6-OCH ₃ ,8,4'-NH ₂	3.43	3.6709	0.0683
39	7,8-OH,3',4',5'-OCH ₃	3.40	4.0058	0.1512
40	3-COOCH ₃ ,4'-OH	3.36	3.7081	0.0939
41	4'-OH	3.30	3.7081	0.1101
42	7-OH,6,3'-NH ₂	3.30	3.3419	0.0125
43	7-OH,6,8,4'-NH ₂	3.12	3.3419	0.0664
44	3-COOCH ₃ ,4'-NH ₂	3.09	3.3419	0.0754
45	3-COOH,7-OCH ₃ ,4'-OH	2.99	3.3262	0.1011
46	7,4'-OH,3',5'-OCH ₃	2.90	3.3262	0.1281
47	7-OH,6,8,4'-NO ₂	2.81	3.0674	0.0839
48	3-COOH,4'-OH	2.80	3.0674	0.0872
49	5-OCH ₃ ,8,4'-NH ₂	2.79	3.0674	0.0904
50	7-OH,8,4'-NO ₂	2.73	3.3262	0.1793

^a pIC₅₀ = -log (IC₅₀), ^b REP = Relative Error Prediction

2.1. MLR analysis

In the first step, separate stepwise selection-based MLR analyses were performed using different types of descriptors, and then, an MLR equation was obtained utilizing the pool of all calculated descriptors. The results are summarized in Table 3. Correlation coefficient (r^2) matrix for the descriptors used in different MLR equations is shown in Table 4. Collinear descriptors degrade the performance of MLR equations and such models have lowered prediction ability.

In Table 3 the QSAR models derived for different derivatives by using different sets of molecular descriptors are listed. Table 3 provides the resulted equations for the studied compounds. The first equation of Table 3 was found by using chemical descriptors (E₁). This equation explained the negative effect of hydration energy and molecular weight (Mass) of molecules on protein tyrosine kinase inhibitory activity. Equation E₂ shows that among quantum descriptors, most positive charge (MPC) has a negative effect on protein tyrosine kinase inhibitory activity and reveals the presence of columbic interactions between the ligands and receptors. The negative sign of the coefficient of MPC demonstrates that ligands with the least MPC could interact with receptor more efficiently. This indicates that there is probably a negative region in receptor which produces columbic interactions with ligand. Equation E₃ of Table 3 demonstrates the effect of constitutional descriptors. It includes the negative effects of average molecular weight (AMW), number of multiple bonds (nBM) and number of aromatic bonds (nAB) on protein tyrosine kinase inhibitory activity. Molecules with lower coefficient of AMW show better protein tyrosine kinase inhibitory activity and decreasing the number of multiple bonds of compounds results in activity enhancement. The MLR equation of Table 3 was obtained from the pool of topological descriptors (E₄) explained the positive effect of mean information content on the distance equality (ICR), path/walk 4-randic shape index (PW4), average connectivity index chi-4 (X4v) and the negative effect of mean information content vertex degree magnitude (IVDM) and average valence connectivity index chi-1 (X1v) on protein tyrosine kinase

inhibitory activity. This equation describes the structure-activity relationship better than those obtained from the chemical, quantum, constitutional descriptors.

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 rotatable 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 (DM _t), Molecular dipole moment at X-direction (DM _x), Molecular dipole moment at Y-direction (DM _y), Molecular dipole moment at Z-direction (DM _z), Electronegativity ($\chi = -0.5$ (HOMO-LUMO)), Electrophilicity ($\omega = \chi^2/2 \eta$) , Hardness ($\eta = 0.5$ (HOMO+LUMO)), Softness ($S = 1/\eta$).
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).
Substituent electronic descriptors	RMSQ (Root mean square error of charges), SPQ (Sum of positive charges), SNQ (Sum of negative charges), RMSDM (Root mean square of dipole moments at any Cartesian coordinate direction), TDM (Total dipole moment), FRMS (Root mean square force that any atom in constituent molecule see right before the optimization), FMAX (Maximum force on molecule), HOMO (Highest occupied molecular orbital), LUMO (Lowest unoccupied molecular orbital), HD (Hardness), SOF (Softness), EPH (Electrophilicity), EN (Electronegativity).

Table 3. The results of MLR analysis with different types of descriptors.

No.	Descriptor source	MLR Equations	N	R ²	SE	RMS _{CV}	Q ²	F
E ₁	Chemical	pIC ₅₀ = 4.893 (± 0.735) - 0.056 (± 0.017) HE -0.007 (± 0.003) Mass	50	0.40	0.55	0.58	0.32	13.82
E ₂	Quantum	pIC ₅₀ = 6.362 (± 0.565) - 6.805 (± 1.505) MPC	50	0.43	0.53	0.54	0.38	17.44
E ₃	Constitutional	pIC ₅₀ = 3.139 (± 1.250) - 0.438 (± 0.100) nBM - 0.506 (± 0.205) AMW - 0.584 (± 0.266) nAB	50	0.49	0.49	0.51	0.42	19.65
E ₄	Topological	pIC ₅₀ = 17.242 (± 0.605) - 3.374 (± 0.545) IVDM - 53.95 (± 12.355) X1Av + 2.349 (± 0.696) ICR + 24.874 (± 9.569) PW4 + 73.575 (± 33.719) X4A	50	0.72	0.38	0.48	0.58	30.13
E ₅	Geometrical	pIC ₅₀ = -15.093 (± 3.339) + 19.450 (± 3.406) SPH - 0.010 (± 0.002) G(N...O)	50	0.60	0.43	0.47	0.49	17.23
E ₆	Functional group	pIC ₅₀ = 3.672 (± 0.123) - 0.414 (± 0.130) nNO ₂ - 1.098 (± 0.369) nOH _t + 0.160 (± 0.058) nOH	50	0.53	0.45	0.50	0.45	12.67
E ₇	Hansch	pIC ₅₀ = 4.219 (± 0.289) - 0.615 (± 0.202) π ₅ + 1.462 (± 0.555) ΣR' ₃ - 1.379 (± 0.490) ΣR ₈ - 0.249 (± 0.111) L ₃	50	0.53	0.45	0.50	0.45	12.67
E ₈	SED	pIC ₅₀ = -0.708 (± 1.228) - 9.570 (± 2.500) HOMO _{A3} + 1.092 (± 0.308) SNQ8	50	0.82	0.32	0.30	0.61	51.43
E ₉	Molecular descriptor	pIC ₅₀ = -19.763 (± 4.304) - 4.785 (± 1.275) MPC + 25.113 (± 4.142) SPH + 0.849 (± 0.264) SNQ8 - 0.357 (± 0.136) L ₃	50	0.83	0.31	0.28	0.62	52.43

The equation obtained from the effect of geometrical parameter on protein tyrosine kinase inhibitory activity of the studied compounds has been described as E₅ of Table 3. It explains the positive effect of sphericity (SPH) and negative effect of sum of geometrical distances between N...O, *i.e.* G (N...O) on protein tyrosine kinase inhibitory activity. The effect of functional groups on protein tyrosine kinase inhibitory activity of the studied compounds has been described by equation E₆ of Table 3. This three-parametric equation does not have a high statistical quality, which suggests that the protein tyrosine kinase inhibitory activity of the studied molecules is not highly dependent on the type of functional group; but it is dependent on the structural changes induced by variations in functional groups. The negative sign of nNO₂ and nOH_t indicates that molecules with lower number of nitro groups (aliphatic) and tertiary alcohols (aliphatic) bind to protein kinase stronger. On the other hand, number of hydroxyl groups (nOH) represents direct effect on the inhibitory activity of the compounds. The Hansch equation (E₇) shows the importance of steric, electronic and lipophilic factors on protein tyrosine kinase inhibitory activity. These factors are described by L₃ (Length parameter of C₃ substituent), ΣR'₃, ΣR₈ (Swain and Lupton field parameter of C-R₃ and C-R₈ substitutes) and π₅ (lipophilic parameter of C₅ substitute), respectively. The negative coefficient of π₅ indicates that lipophilic substituents at R₅ are not favorable for binding affinity. This equation shows the positive

effect of $\Sigma R'_3$ and the negative effect of ΣR_8 on the inhibitory activity of the compounds. In addition the negative effect of L_3 describes that the presence of bulky groups at C_3 leads to decreased activity because bulky groups hinder strong interaction between ligands and the enzyme. The SED equation (E_8) shows the importance of SED factors on protein tyrosine kinase inhibitory activity. One of the parameters is molecular orbital energy $HOMO_{A3}$ (Highest occupied molecular orbital parameter of C_3 substitute) and the other one is $SNQ8$ (Sum of negative charges parameter of C_8 substitute). It explains the positive effect of $HOMO_{A3}$ and negative effect of $SNQ8$ on protein tyrosine kinase inhibitory activity.

The last Equation (E_9) was obtained from the all types of calculated descriptors. Stepwise selection and elimination of variables produced a four-parametric QSAR equation. This equation shows that geometrical (SPH), quantum (MPC), Hansch (L_3) and SED ($SNQ8$) parameters are major factors that affect protein tyrosine kinase inhibitory activity of compounds. Among these descriptors MPC and L_3 have negative effects and the others have positive effects on the protein tyrosine kinase inhibitory activity.

2.2. Free-Wilson analysis

The simple Free-Wilson analysis (FWA) was considered to indicate which substituents on ring **B** and chromone moiety contribute to protein tyrosine kinase inhibitory activity and which ones detract from activity [33]. As indicated in Table 1, the molecules used in this study have a phenyl ring (ring **B**) and chromone moiety with different types of substituents in different positions of the ring. Some important substituents such as methoxyl, hydroxyl and amine are used in calculations. Therefore, the descriptors data matrix built for the FWA has 44 rows (i.e., number of selected molecules for FWA) and 24 columns (i.e., three substituents at eight substitution positions on the flavonoid structure). The elements of the descriptor data matrix are 1 or 0, to indicate the presence or absence of a given substituent in a specified position in a molecule, respectively. The following two-parametric equation was found between the activity data (y) and the Free-Wilson type descriptors data matrix:

$$pIC_{50} = 3.893 (\pm 0.089) + 0.439 (\pm 0.207) R'_3\text{-Hydroxyl} - 1.103 (\pm 0.534) R_5\text{-Methoxyl} \quad (1)$$

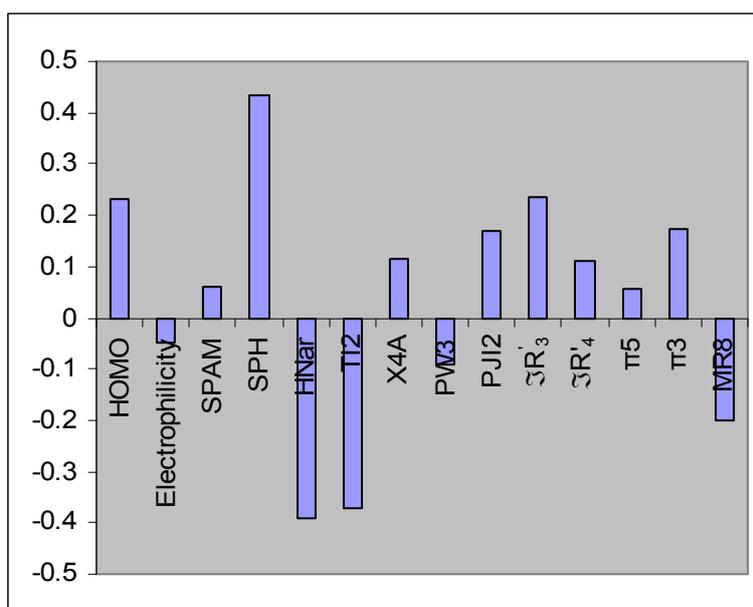
$$R^2 = 0.70, \quad N = 44, \quad F = 31.45, \quad SE = 0.30$$

Equation (1) describes that protein tyrosine kinase inhibitory activity of studied compounds is directly affected by the presence of electron-donating hydroxyl group in the *meta* position (R'_3) of the phenyl ring and most probably this part of the flavonoid molecule interacts with the catalytic domain of the enzyme. The same result was obtained by other researchers [27]. A methoxyl group on $C-R_5$ detracts from the inhibitory activity, according to this equation.

2.3. GA-PLS analysis

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 multicollinearity problem in the descriptors is omitted by PLS analysis. Because a minimal number of latent variables are 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. To do so, many different GA-PLS runs were conducted using different initial set of populations. The data set ($n = 50$) was divided into two group: calibration set ($n = 40$) and prediction set ($n = 10$). Given 40 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 14 indices, four of them being those obtained by MLR. The PLS estimate of coefficients for these descriptors are given in Figure 1. As it observed, a combination of quantum, topological, geometrical and Hansch descriptors have been selected by GA-PLS to account the protein tyrosine kinase inhibitory activity of flavonoid derivatives. The majority of these descriptors are topological indices. The resulted GA-PLS model possessed a high statistical quality $R^2 = 0.74$ and $Q^2 = 0.61$. The predictive ability of the model was measured by applying to 10 external test set molecules. The squared correlation coefficient for prediction was 0.82 and standard error of prediction was 0.30. The values of pIC_{50} using GA-PLS model (refined from cross-validation or external prediction set) along with the corresponding relative errors of prediction (REP) are shown in Table 1. Very small values of relative errors (between ± 0.40) confirm the accuracy of the proposed GA-PLS model for modeling protein tyrosine kinase inhibitory activity of the studied flavonoid derivatives.

Figure 1. PLS regression coefficients for the variables used in GA-PLS model.

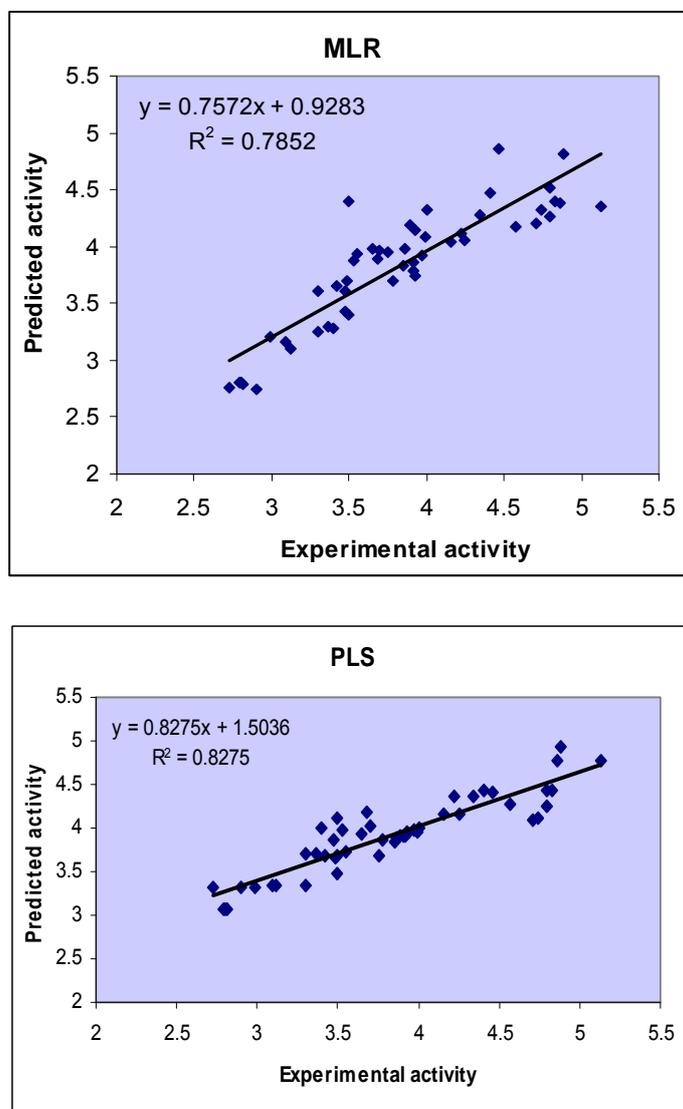


Comparison between the results obtained by GA-PLS and MLR methods indicates higher accuracy of GA-PLS method in describing the inhibitory activity of flavonoid derivatives toward protein

tyrosine kinase enzyme. The difference in accuracy of the two regression methods used in this study is visualized in Figure 2 by plotting the predicted activity (by cross-validation) against the experimental values. Obviously, two linear models represented scattering of data around a straight line with slope close to one. As it is observed, the plot of data resulted by GA-PLS represents the lowest scattering and the plot obtained by MLR analysis (which is obtained from E_9) is in the second order of accuracy.

To measure the significance of the 14 selected PLS descriptors in the protein tyrosine kinase inhibitory activity; VIP was calculated for each descriptor [34]. The VIP analysis of PLS equation is shown in Figure 3. VIP shows that HNar and TI2, which are topological, and SPH which is a geometrical parameter, are the most important indices in the QSAR equation derived by PLS analysis. In addition, quantum parameters such as (HOMO) and Hansch ($\sigma R'_3$) have been found to be moderately influential parameters.

Figure 2. Plots of the cross-validated predicted activity against the experimental activity for the QSAR models obtained by MLR, GA-PLS methods.

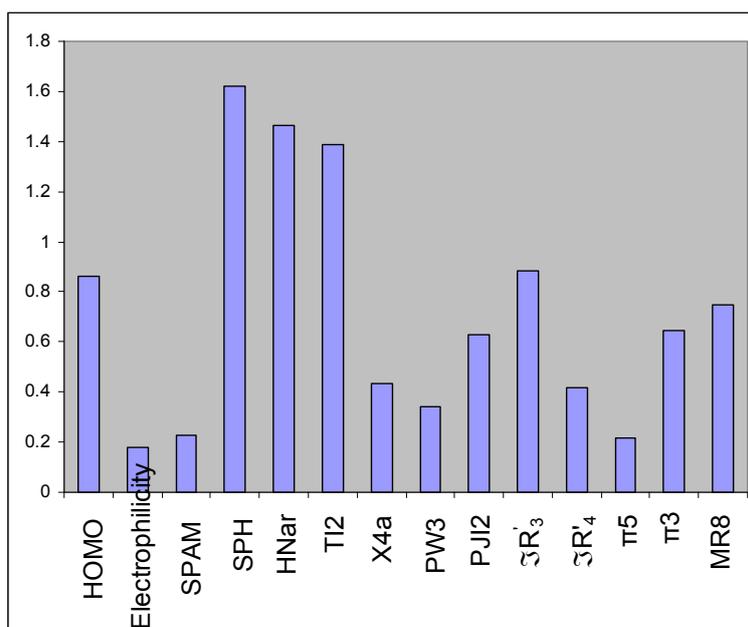


3. Methodology

3.1. Software

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 the Polak-Ribiere algorithm until the root mean square gradient was $0.01 \text{ kcal mol}^{-1}$. The resulted geometry was transferred into Dragon program package, which was developed by Milano Chemometrics and QSAR Group [35]. The z-matrix of the structures was provided by the software and transferred to the 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 performed using Gaussian 98 program [36].

Figure 3. Plot of variables important in projection (VIP) for the descriptors used in GA-PLS model.



3.2. Activity data & descriptor generation

The biological data used in this study are protein tyrosine kinase inhibitory activity, $-\log(\text{IC}_{50})$, of a set of 50 flavonoid analogues [32]. The structural features and biological activity of these compounds are listed in Table 1 and then used for subsequent QSAR analysis as dependent variables. The large number of molecular descriptors was calculated using Hyperchem, Dragon package and Gaussian 98. Some chemical parameters including molecular volume (V), molecular surface area (SA), hydrophobicity (Log P), hydration energy (HE) and molecular polarizability (MP) were calculated using Hyperchem Software. The Dragon software calculated different functional groups, topological, geometrical and constitutional descriptors for each molecule. Gaussian 98 was employed for

calculation of 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.* [37]. Classical substituent constants including hydrophobic constant (π), the Hammett electronic constants (σ), the Taft field effect (FI), resonance (R) substituent and steric (molar refractivity MR and STERIMOL) constants were also used as descriptor in this study [38]. The calculated descriptors for each molecule are summarized in Table 2.

3.3. Data screening & model building

The selected descriptors from each class and the experimental data were analyzed by the stepwise regression SPSS (version 12.0) software. The calculated descriptors were collected in a data matrix whose number of rows and columns were the number of molecules and descriptors, respectively. Multiple linear regression (MLR) and partial least squares (PLS) were used to derive the QSAR equations and feature selection was performed by the use of genetic algorithm (GA). 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 and therefore we used genetic algorithm (GA-PLS) to select the best variables.

Application of PLS allows the construction of larger QSAR equations, while still avoiding overfitting and eliminating most variables. PLS is normally used in combination with cross-validation to obtain the optimum number of components [39, 40]. 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 [41].

3.4. 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, variable important in projection (VIP) was employed [34]. VIP values reflect the importance of terms in 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 [8].

3.5. Substituent electronic descriptors (SED)

Electronic descriptors obtained from quantum chemical calculations have found major popularity and there is a challenge between calculation complexity and accuracy to select the quantum chemical calculation methods (i.e., semi-empirical and *ab initio*) [42]. To simplify the quantum chemical calculations Hemmateenejad *et al.* recently have hypothesized that the calculations could be performed on the substituents instead of whole molecular structures and the resulting electronic features can be considered as electronic descriptors which have found major popularity in QSAR/QSPR studies

[43,44]. Hemmateenejad *et al.* proposed substituent electronic descriptors (SED) as an alternative to both substituent constants and molecular descriptors [43]. SED analysis for each substituent was used in our study and the calculated descriptors are listed in Table 2. They can be classified into three different electronic categories including local charges, dipoles and orbital energies. Since most of the constituents are open shell quantum species (due to being in doublet quantum state as a radical molecule), a difference in energy between two electronic energy populations, alpha (spine up) and beta (spine down) can be seen using Gaussian 98. It provides some additional descriptors HOMO_A, HOMO_B, LUMO_A, LUMO_B, HAD, HD_B, SOF_A, SOF_B, EN_A, EN_B, EPH_A, and EPH_B stem from two different alpha and beta electronic population energy, where the subscript A and B stand for alpha and beta population of electronic energy, respectively. Therefore, a total of 26 electronic descriptors were calculated for each substituent.

4. Conclusions

Quantitative relationships between molecular structure and protein tyrosine kinase inhibitory activity of flavonoid derivatives were discovered by two chemometrics methods: MLR and GA-PLS. Different QSAR models revealed that SED parameters have significant impact on protein tyrosine kinase inhibitory activity of the compounds. In this series a significant role of topological and geometrical parameters on the inhibitory activity was observed. Using the pool of all types of calculated descriptors a new QSAR model was derived for these compounds. In this model the importance of quantum, geometrical, SED and Hansch parameters have an effect on protein tyrosine kinase inhibitory activity was indicated. A comparison between the two statistical methods employed indicated that GA-PLS represented superior results. The resulted GA-PLS model possessed a high statistical quality ($R^2 = 0.74$ and $Q^2 = 0.61$) for predicting the activity of the inhibitors. The models proposed in present work are more useful in describing QSAR of flavonoid derivatives as p56^{lck} protein tyrosin kinase Inhibitors than those proposed previously.

Acknowledgments

This work was supported by Isfahan Pharmaceutical Sciences Research Center. The authors wish to thank Dr. Bahram Hemmateenejad for his advice on various aspects of this research.

References

1. Hansch, C.; Hoekman D.; Gao, H. Comparative QSAR: Toward a Deeper Understanding of Chemicobiological Interactions. *Chem. Rev.* **1996**, *96*, 1045-1076.
2. Hansch, C.; Maloney, P.P.; Fujita, T.; Muir, R.M. Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. *Nature* **1962**, *194*, 178-180.
3. Hemmateenejad, B. Correlation Ranking Procedure for Factor Selection in PC-ANN Modeling and Application to ADMETox Evaluation. *Chemom. Intell. Lab. Syst.* **2005**, *75*, 231-245.

4. Fujita, T.; Iwasa, J.; Hansch, C. A New Substituent Constant, π , Derived from Partition Coefficients. *J. Am. Chem. Soc.* **1964**, *86*, 5175-5180.
5. Hansch, C. Quantitative Approach to Biochemical Structure-Activity Relationships. *Acc. Chem. Res.* **1968**, *2*, 232-239.
6. Hansch, C.; Clayton, J.M. Lipophilic Character and Biological Activity of Drugs II: The Parabolic Case. *J. Pharm. Sci.* **1973**, *62*, 1-21.
7. Agatonovic-Kustrin, S.; Tucker, I.G.; Zecevic, M.; Ziva-novic, L.J. Prediction of Drug Transfer into Human Milk from Theoretically Derived Descriptors. *Anal. Chem. Acta* **2000**, *418*, 181-195.
8. Mohajeri, A.; Hemmateenejad, B.; Mehdipour A.; Miri, R. Modeling Calcium Channel Antagonistic Activity of Dihydropyridine Derivatives Using QTMS Indices Analyzed by GA-PLS and PC-GA-PLS. *J. Mol. Graph. Model.* **2008**, *26*, 1057-1065.
9. Ullrich A.; Schlessinger, J. Signal Transduction by Receptors with Tyrosine Kinase Activity. *Cell*, **1990**, *61*, 203-212.
10. Bishop, J.M. The Molecular Genetics of Cancer. *Science* **1987**, *235*, 305-311.
11. Blume-Jensen, P.; Hunter, T. Oncogenic Kinase C Signalling. *Nature* **2001**, *411*, 355-365.
12. Hunter, T. Signaling — 2000 and Beyond. *Cell* **2000**, *100*, 113-127.
13. Schlessinger, J. Cell Signaling by Receptor Tyrosine Kinases. *Cell* **2000**, *103*, 211-225.
14. Hanahan, D.; Weinberg, R.A. The Hallmarks of Cancer. *Cell* **2000**, *100*, 57-70.
15. Cantley, L.C.; Auger, K.R.; Carpenter, C.; Duckworth, B.; Graziani, A.; Kapeller, R.; Oloff, S. Oncogenes and Signal Transduction. *Cell* **1991**, *64*, 281-302.
16. Groundwater, P.W.; Solomons, K.R.H.; Drewe J.A.; Munawar, M.A. *Progress in Medicinal Chemistry*; Ellis, G.P., Luscombe, D.K., Eds.; Elsevier Science B.V.: Amsterdam, 1996; pp. 233-329.
17. Bolen, J.B.; Veillette, A.; Schwartz, A.M.; DeSeau, V.; Rosen, N. Activation of pp60c-src Protein Kinase Activity in Human Colon Carcinoma. *Proc. Natl. Acad. Sci. USA* **1987**, *84*, 2251-2255.
18. Slamon, D.J.; Clark, G.M.; Wong, S.G.; Levin, W.J.; Ullrich A.; McGuire, W.L. Human Breast Cancer: Correlation of Relapse and Survival with Amplification of the HER-2/neu Oncogene. *Science* **1987**, *235*, 177-182.
19. Yamamoto, T.; Kamata, N.; Kawano, H.; Shimizu, S.; Kuroki, T.; Toyoshima, K.; Rikimaru, K.; Nomura, N.; Ishizaki, R.; Pastan, I.; Gamou S.; Shimizu, N. High Incidence of Amplification of the Epidermal Growth Factor Receptor Gene in Human Squamous Carcinoma Cell Lines. *Cancer Res.* **1986**, *46*, 414-416.
20. Weil, R.; Veillette, A. Signal Transduction by the Lymphocyte-Specific Tyrosine Protein Kinase p56^{lck}. *Current Topics Micro. Immunol.* **1996**, *205*, 63-87.
21. Anderson, S.J.; Levin, S.D.; Perlmutter, R.M. Involvement of the Protein Tyrosine Kinase p56^{lck} in T Cell Signaling and Thymocyte Development. *Adv. Immunol.* **1994**, *56*, 151-178.
22. Bishop, J.M. Cellular Oncogenes and Retroviruses. *Annu. Rev. Biochem.* **1983**, *52*, 301-354.
23. Cushman, M.; Nagarathnam, D.; Burg, D.L.; Geahlen, R.L. Synthesis and Protein-Tyrosine Kinase Inhibitory Activities of Flavonoid Analogues. *J. Med. Chem.* **1991**, *34*, 798-806.
24. Cushman, M.; Zhu, H.; Geahlen R.L.; Kraker, A.J. Synthesis and Biochemical Evaluation of a Series of Aminoflavones as Potential Inhibitors of Protein-Tyrosine Kinases p56^{lck}, EGFr, and p60v-src *J. Med. Chem.* **1994**, *37*, 3353-3362.

25. Bylka, W.; Matlawska, I.; Pilewski, N.A. Natural Flavonoids as Antimicrobial Agents. *JANA* **2004**, *7*, 24-31.
26. Thakur, A.; Vishwakarma, S.; Thakur, M. QSAR Study of Flavonoid Derivatives as p56^{lck} Tyrosine Kinase Inhibitors. *Bioorg. Med. Chem.* **2004**, *12*, 1209-1214.
27. Nikolovska-Coleska, Ž.; Suturkova, L.; Dorevski, K.; Krbavcic, A.; Solmajer, T. Quantitative Structure-Activity Relationship of Flavonoid Inhibitors of p56^{lck} Protein Tyrosine Kinase: A Classical/Quantum Chemical Approach. *Quant. Struct.-Act. Relat.* **1998**, *17*, 7-13.
28. Novic, M.; Nikolovska-Coleska, Ž.; Šolmajer, T. Quantitative Structure-Activity Relationship of Flavonoid p56^{lck} Protein Tyrosine Kinase Inhibitors. A Neural Network Approach. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 990-998.
29. Oblak, M.; Randic, M.; Solmajer, T. Quantitative Structure-Activity Relationship of Flavonoid Analogues.3. Inhibition of p56^{lck} Protein Tyrosine Kinase. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 994-1001.
30. Stefanic-Petek, A.; Krbavcic, A.; Solmajer, T. QSAR of Flavonoids: 4. Differential Inhibition of Aldose Reductase and p56^{lck} Protein Tyrosine Kinase *Croatica Chemica Acta* **2002**, *75*, 517-529.
31. Meyer, M. *Ab initio* Study of Flavonoid. *Int. J. Quantum Chem.* **2000**, *76*, 724-732.
32. Deeb, O.; Clare, B.W. QSAR of Aromatic Substances: Protein Tyrosin Kinase Inhibitory Activity of Flavonoid Analogues. *Chem. Biol. Drug Des.* **2007**, *70*, 437-449.
33. Free, S.M.J.R.; Wilson, J.W. A Mathematical Contribution to Structure-Activity Studies. *J. Med. Chem.* **1964**, *7*, 395-399.
34. Olah, M.; Bologna, C.; Oprea, T.I. An Automated PLS Search for Biologically Relevant QSAR Descriptors. *J. Comput. Aided Mol. Des.* **2004**, *18*, 437-449.
35. Todeschini, R. Milano Chemometrics and QSPR Group. <http://michem.disat.unimib.it/>, accessed 9 September, 2008.
36. Frisch, M.J.; Trucks, M.J.; Schlegel, H.B.; Scuseria, G.E.; Robb, M.A.; Cheeseman, J.R.; Zakrzewski, V.G.; Montgomery, J.A.; Stratmann, J.R.; Burant, J.C. *et al.* Gaussian 98, Revision A.7, Gaussian, Inc.: Pittsburgh, PA, 1998.
37. Roy, K. QSAR of Adenosine Receptor Antagonists II: Exploring Physicochemical Requirements for Selective Binding of 2-arylpyrazolo [3,4-c]quinoline Derivatives with Adenosine A¹ and A³ Receptor Subtypes *QSAR. Comb. Sci.* **2003**, *22*, 614-621.
38. Hansch, C.; Leo, A.; Taft, R.W. A Survey of Hammett Substituent Constants and Resonance and Field Parameters. *Chem. Rev.* **1991**, *91*, 165-195.
39. 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*, 3737-3743.
40. Leardi, R. Genetic Algorithms in Chemometrics and Chemistry: A Review. *J. Chemometrics.* **2001**, *15*, 559-569.
41. Hemmateenejad, B. Optimal QSAR Analysis of the Carcinogenic Activity of Drugs by Correlation Ranking and Genetic Algorithm-Based. *J. Chemometrics.* **2004**, *18*, 475-485.
42. Wang, J.; Zhang, L.; Yang, G.; Zhan, C.G. 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**, *44*, 2099-2105.

43. Hemmateenejad B.; Sanchooli, M. Substituent Electronic Descriptors for Fast QSAR/QSPR. *J. Chemometrics.* **2007**, *21*, 96-107.
44. Smeyers, Y.G.; Bouniam, L.; Smeyers, N.J.; Ezzamarty, A.; Hernandez-Laguna, A.; Sainz-Diaz, C.I. Quantum Mechanical and QSAR Study of Some α -Arylpropionic Acids as Anti-Inflammatory Agents. *Eur. J. Med. Chem.* **1998**, *33*, 103-112.

© 2008 by the authors; licensee Molecular Diversity Preservation International, Basel, Switzerland. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).