Available online at www.pharmaresearchlibrary.com/ijcps ISSN: 2321-3132

Research Article

International Journal of Chemistry and Pharmaceutical Sciences
IJCPS, 2013: Vol.1(5): 319-330
www.pharmaresearchlibrary.com/ijcps

A Quantitative Structure-Activity Relationship Study of Caffeic Acid Amides as Selective Inhibitors of Matrix Metalloproteinase

Prithvi Singh*
Department of Chemistry, S.K. Government Post-Graduate College, Sikar-332 001, India
*E-mail: psingh_sikar@rediffmail.com

Abstract

A quantitative structure-activity relationship (QSAR) study is carried out on caffeic acid amides as the inhibitors of matrix metalloproteinase (MMP). The inhibition action, for the sub-type MMP-9, is quantitatively analyzed following the non-parametric (Fujita-Ban) and parametric approaches. The Fujita-Ban analysis has identified the substituents which imparted highest contributions to parent moiety, makes it feasible to design more active analogues of the series. The parametric approach, on the other hand, utilized molecular 2D-descriptors to develop statistical validated models through combinatorial protocol in multiple linear regression analysis (CP-MLR). The descriptors, participated in the most significant models, have highlighted the role of Moran autocorrelations of lag-4 and lag-6 weighted, respectively, by atomic masses (MATS4m) and atomic Sanderson electro negativities (MATS6e). Additionally, the Balaban-type index obtained from polarizability weighted distance matrix (Jhetp) and Randic shape index representing the ratio of path over walk count of order 4 (PW4) also showed prevalence in the rationalization of activity profiles. The partial least squares (PLS) analysis further confirmed the dominance of the CP-MLR identified descriptors. The guidelines delineated by the Fujita-Ban approach, statistically validated models and PLS analysis, facilitated in exploring some potential analogues of the series. Applicability domain (AD) analysis revealed that the suggested models have acceptable predictability.

Keywords: Caffeic acid amide derivatives, MMP-9 inhibitors, combinatorial protocol in multiple linear regression (CP-MLR) analysis, QSAR study, DRAGON 2D-descriptors.

Introduction

Matrix metalloproteinases (MMPs) represent a family of zinc dependent endopeptidases which is involved in the breakdown of components of the extracellular matrix and helps in tissue remodeling [1]. The process is vital in embryonic development, pregnancy, growth and wound healing. Usually the activity of MMPs is restricted by the equilibrium between synthesis of active MMPs and the presence of endogenous inhibitors, namely the tissue inhibitor of metalloproteinases (TIMPs) [1]. During tumor progression this equilibrium is disturbed and the increased level of certain MMPs is thought to be involved in metastatic tumor dispersion and angiogenesis leading to a malignant phenotype [2,3]. The zinc dependent endopeptidases are then secreted as inactive zymogens in the extracellular matrix by the tumor cells or by neighboring stromal cells close to tumor cells. The MMP-2 (gelatinase A) and MMP-9 (gelatinase B) appear to play significant role in these progressions, considering that they are engaged in metastatic tumor dispersion and angiogenesis [4,5].

From X-ray crystallography, NMR analysis and homology modeling, the MMPs have been classified into two broad structural classes. Those with a relatively deep S1′ pocket includes MMP-2, -3, -8, -9, and -13 and those with a shallow S1′ pocket incorporates MMP-1 and MMP-7 [6,7]. Therefore, inclusion of an extended P1′ group leads to selective inhibition, whereas the presence of smaller P1′ groups leads to broad spectrum inhibition. These results have helped to develop more selective second generation inhibitors against the specific MMPs. The MMPIs that may be obtained from natural resources such as herbs, plants, fruits, and other agriculture products have drawn considerable attention in recent years [8]. Caffeic acid [CA; 3-(3,4-dihydroxyphenyl)acrylic acid] which was found in fruits, vegetables, wine, olive oil and coffee [9], has been shown to inhibit the activity of MMP-9 while caffeic

acid phenethyl ester (CAPE) which was extracted from honeybee propolis [10] and has been synthesized by esterification of CA [11], could selectively inhibited MMP-2 and -9 but not -1, -3, -7 [12]. However, the CAPEs have limited use due to their metabolically unstable ester groups [13-15]. Nonetheless, many modified caffeic acid amides have been reported to reveal stable behavior [16].

In view of this, Shi et al. have recently reported [17] a congeneric series of caffeic acid amides with extended P1′ group to investigate their selective inhibition activities against MMP-2 and MMP-9. They have also investigated the MMPs inhibitory effects of the substitution on the benzene ring of CA. For this, three acids, namely, (E)-3-(3-hydroxyphenyl)acrylic acid, (E)-3-(4-hydroxyphenyl)acrylic acid and cinnamic acid have been amidated with extended P1′ group and analyzed for their inhibition actions on MMP-1, MMP-2 and MMP-9. The reported study describing structure-activity relationships (SARs) was, however, targeted at the alterations of substituents at different positions and provided no rationale to reduce the trial-and-error issues. Hence in the present communication, a 2D-quantitative SAR (2D-QSAR) study is conducted on these analogues to provide the rationale for drug-design and to explore the structural requirements for possible interaction. In the congeneric series, where a relative study is being carried out, the 2D-descriptors may play important role to quantify the activity profiles of the compounds. The novelty and importance of a 2D-QSAR study is due to its simplicity for the calculations of different descriptors and their interpretation (in physical sense) to explain the biological activity of compounds at molecular level.

Materials and Methods

Data set

In the present work, the compounds (general structure in Figure 1) and their inhibition activities, IC_{50}s, were taken from the literature [17]. The IC_{50} value represents the concentration

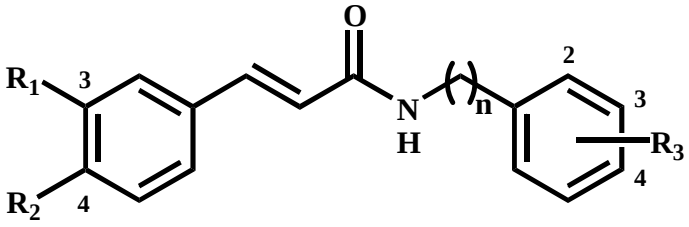

Figure.1 General structure of caffeic acid amide derivatives.

of a compound needed to bring out 50% inhibition of MMP-9 and the same is expressed as pIC_{50} (= $-\log IC_{50}$) on a molar basis. The compounds and their inhibition activities, pIC_{50} values are listed in Table 1.

Table.1 Observed and modeled MMP-9 inhibition activity of caffeic acid amides

S.No.	R_1	R_2	R_3	n	pIC_{50} (M)				
					Obsd.[a]	Calcd.			
						F.B.	Eq. (6)	Eq. (7)	PLS
1[b]	OH	OH	H	0	8.15	8.18	7.62	7.59	7.81
2	OH	OH	H	1	8.14	8.18	8.22	8.03	8.24
3[b]	OH	OH	H	2	8.14	8.18	8.06	7.82	7.91
4[b]	OH	OH	2-OH	0	8.28	8.29	8.07	7.99	8.11
5	OH	OH	3-OH	0	5.78	5.85	6.01	5.93	6.04
6	OH	OH	4-OH	0	8.63	8.48	8.13	8.07	8.05
7	OH	OH	2-OMe	0	8.25	8.28	7.77	7.72	7.72
8[b]	OH	OH	3-OMe	0	5.79	5.83	7.01	6.87	6.87
9[b]	OH	OH	4-OMe	0	8.48	8.44	8.34	8.08	8.38
10	OH	OH	2-Me	0	8.23	8.27	8.02	7.77	7.76
11	OH	OH	3-Me	0	5.76	5.81	6.60	6.55	6.13
12[b]	OH	OH	4-Me	0	8.44	8.43	7.98	7.86	7.80
13	OH	OH	2-F	0	8.11	8.12	8.03	8.02	8.03
14	OH	OH	3-F	0	5.73	5.81	6.04	6.00	6.09
15	OH	OH	4-F	0	8.14	8.14	8.07	8.03	7.86
16	OH	OH	2-Cl	0	8.09	8.08	7.90	7.59	8.33
17	OH	OH	3-Cl	0	5.72	5.79	6.09	6.01	6.33
18	OH	OH	4-Cl	0	8.10	8.10	7.80	7.65	7.90
19	OH	H	H	0	8.12	8.13	7.46	7.76	7.59
20[b]	OH	H	H	1	8.12	8.13	8.09	8.20	8.07

21	OH	H	H	2	8.12	8.13	7.90	7.95	7.68
22[b]	OH	H	2-OH	0	8.24	8.24	7.95	8.17	7.94
23	OH	H	3-OH	0	5.81	5.80	5.40	5.59	5.41
24	OH	H	4-OH	0	8.47	8.43	7.97	8.24	7.82
25	OH	H	2-OMe	0	8.24	8.23	7.72	7.94	7.64
26	OH	H	3-OMe	0	5.78	5.77	6.63	6.71	6.48
27	OH	H	4-OMe	0	8.47	8.39	8.25	8.29	8.24
28	OH	H	2-Me	0	8.23	8.21	8.03	8.09	7.73
29	OH	H	3-Me	0	5.75	5.76	6.34	6.57	5.85
30	OH	H	4-Me	0	8.45	8.37	7.90	8.11	7.68
31[b]	OH	H	2-F	0	8.10	8.07	7.85	8.12	7.79
32[b]	OH	H	3-F	0	5.73	5.75	5.48	5.68	5.52
33	OH	H	4-F	0	8.11	8.09	7.87	8.16	7.58
34	OH	H	2-Cl	0	8.08	8.03	7.71	7.69	8.12
35[b]	OH	H	3-Cl	0	5.71	5.74	5.51	5.63	5.77
36[b]	OH	H	4-Cl	0	8.09	8.04	7.61	7.79	7.65
37[b]	H	OH	H	0	8.11	8.11	7.84	7.90	8.01
38	H	OH	H	1	8.11	8.10	8.40	8.25	8.42
39	H	OH	H	2	8.11	8.10	8.12	7.95	7.95
40	H	OH	2-OH	0	8.24	8.21	8.32	8.33	8.35
41	H	OH	3-OH	0	5.80	5.77	5.70	5.64	5.77
42	H	OH	4-OH	0	8.37	8.41	8.34	8.36	8.25
43	H	OH	2-OMe	0	8.23	8.20	7.97	7.97	7.93
44[b]	H	OH	3-OMe	0	5.78	5.75	6.85	6.70	6.75
45	H	OH	4-OMe	0	8.36	8.37	8.52	8.36	8.55
46[b]	H	OH	2-Me	0	8.22	8.19	8.42	8.24	8.16
47	H	OH	3-Me	0	5.77	5.74	6.63	6.65	6.18
48	H	OH	4-Me	0	8.36	8.35	8.30	8.24	8.12
49[b]	H	OH	2-F	0	8.05	8.05	8.20	8.28	8.19
50	H	OH	3-F	0	5.79	5.73	5.73	5.71	5.82
51	H	OH	4-F	0	8.08	8.07	8.22	8.26	7.99
52	H	OH	2-Cl	0	8.03	8.00	8.17	7.91	8.61
53	H	OH	3-Cl	0	5.74	5.72	5.82	5.73	6.12
54	H	OH	4-Cl	0	8.04	8.02	8.05	7.95	8.15
55	H	H	H	0	8.10	8.06	7.82	8.05	8.09
56	H	H	H	1	8.09	8.05	8.42	8.42	8.52
57	H	H	H	2	8.09	8.05	7.96	7.87	7.88
58	H	H	2-OH	0	8.14	8.16	8.29	8.40	8.45
59[b]	H	H	3-OH	0	5.75	5.72	4.85	4.82	5.11
60	H	H	4-OH	0	8.21	8.36	8.27	8.41	8.27
61	H	H	2-OMe	0	8.14	8.15	7.99	8.09	8.06
62	H	H	3-OMe	0	5.70	5.70	6.28	6.17	6.34
63	H	H	4-OMe	0	8.21	8.32	8.51	8.44	8.64
64	H	H	2-Me	0	8.13	8.14	8.74	8.69	8.60
65	H	H	3-Me	0	5.72	5.69	6.34	6.45	6.04
66	H	H	4-Me	0	8.20	8.30	8.47	8.56	8.41
67[b]	H	H	2-F	0	7.98	8.00	8.07	8.22	8.17
68	H	H	3-F	0	5.72	5.68	4.97	4.99	5.22
69	H	H	4-F	0	7.99	8.02	8.11	8.23	7.98
70	H	H	2-Cl	0	7.86	7.95	8.13	7.91	8.72
71	H	H	3-Cl	0	5.74	5.66	4.89	4.81	5.38
72	H	H	4-Cl	0	7.90	7.97	7.97	7.99	8.17

[a]IC_{50} is the concentration of a compound required to bring out 50% inhibition of MMP-9, which is expressed as pIC_{50} (= $-\log IC_{50}$) on a molar basis; taken from Ref. [17]; [b]Test-set compound.

As the reported activity variation for inhibition of MMP-1 was very small, therefore, it was improper to consider the same as dependent variable for present investigation. Also the activity profiles for MMP-2 were perfectly correlated to that of MMP-9 ($r = 0.999$, $s = 0.049$, $F(1,70) = 36239.387$) for 72 data-points of Table 1, suggesting that the behavior of these compounds were similar towards MMP-2 and MMP-9 enzymes. Thus, it is pertinent to analyze the MMP-9 inhibition activity in relation with appropriate descriptors describing molecular structures. For modeling purpose, the data-set was further divided into training-set and test-set. Nearly 25% of the total compounds were

selected for the test-set while remaining compounds were included in the training-set. The statistical significant models developed from the training-set were externally validated using the identified test-set.

Molecular descriptors

The structures of the 72 caffeic acid amides under study were drawn in ChemDraw [18] using the standard procedure. All these structures were ported to DRAGON software [19] for computation of the descriptors corresponding to 0D-, 1D-, and 2D-classes. Table 2 provides the definition and scope of these descriptor classes in addressing the structural features of the compounds. The combinatorial protocol in multiple linear regression (CP-MLR) computational procedure [20] was used for present work in developing 2D-QSAR models.

Table.2 Descriptor classes used for the analysis of caffeic acid amides for modeling MMP-9 inhibition activity

Descriptor class (acronyms)	Definition and scope
Constitutional (CONST)	Dimensionless or 0D descriptors; independent from molecular connectivity and conformations.
Topological (TOPO)	2D-descriptor from molecular graphs and independent conformations.
Molecular walk counts (MWC)	2D-descriptors representing self-returning walk counts of different lengths.
Modified Burden eigenvalues (BCUT)	2D-descriptors representing positive and negative eigenvalues of the adjacency matrix, weights the diagonal elements and atoms.
Galvez topological charge indices (GLVZ)	2D-descriptors representing the first 10 eigenvalues of corrected adjacency matrix.
2D-autocorrelations (2DAUTO)	Molecular descriptors calculated from the molecular graphs by summing the products of atom weights of the terminal atoms of all the paths of the considered path length (the lag).
Functional groups (FUNC)	Molecular descriptors based on the counting of the chemical functional groups.
Atom-centred fragments (ACF)	Molecular descriptors based on the counting of 120 atom-centred fragments, as defined by Ghose-Crippen.
Empirical (EMP)	1D-descriptors represent the counts of non-single bonds, hydrophilic groups and ratio of the number of aromatic bonds and total bonds in an H-depleted molecule.
Properties (PROP)	1D-descriptors representing molecular properties of a molecule.

Model development

The CP-MLR is a 'filter'-based variable selection procedure for model development in QSAR studies [20]. The procedural aspects and implementation of this procedure are discussed in our recent publications [21-26]. The developed computer program, based on CP-MLR procedure, is interfaced with four filters which make the variable selection process efficient and lead to a unique solution. Filter-1 seeds the variables by way of limiting inter-parameter correlations to predefined level (upper limit ≤ 0.79); filter-2 controls the variables entry to a regression equation through t-values of coefficients (threshold value ≥ 2.0); filter-3 provides comparability of equations with different number of variables in terms of square root of adjusted multiple correlation coefficient of regression equation, r-bar; filter-4 estimates the consistency of the equation in terms of cross-validated Q^2 with leave-one-out (LOO) cross-validation as default option (threshold value $0.3 \leq Q^2 \leq 1.0$). In order to collect the descriptors with higher information content and explanatory power, the threshold of filter-3 was successively incremented with increasing number of descriptors (per equation) by considering the r-bar value of the preceding optimum model as the new threshold for next generation.

Model validation

In order to discover any chance correlations associated with the models recognized in CP-MLR, each model has been put to a randomization test [27,28] by repeated randomization of the activity to ascertain the chance correlations, if any, associated with them. For this, every model has been subjected to 100 simulation runs with scrambled activity. The scrambled activity models with regression statistics better than or equal to that of the original activity model have been counted, to express the percent chance correlation of the model under scrutiny. A statistical index, r^2_{randY}(sd), representing the mean random squared multiple correlation coefficient of the regressions in the activity (Y) randomization study with its standard deviation from 100 simulations, has been computed to ensure that none of these scrambled model is superior to that of original model. The internal consistency (validation) of each developed model was ascertained through the cross-validated index, Q^2, from leave-one-out (Q^2_{LOO}) and leave-five-out (Q^2_{L5O}) procedures. A value greater than 0.5 of Q^2-index hints towards a reasonable robust model. The external validation or predictive power of derived model is based on test-set compounds. The statistical index r^2_{Test}, representing the squared correlation coefficient between the observed and predicted data of the test-set, was

also computed for this purpose. A value greater than 0.5 of r^2_{Test} suggests that the model obtained from training-set has a reliable predictive power. Goodness of fit of models was assessed by examining the multiple correlation coefficient (r), the standard deviation (s), the F-ratio between the variances of calculated and observed activities (F). A number of additional statistical parameters such as the Akaike's information criterion, AIC [29,30], the Kubinyi function, FIT [31,32] and the Friedman's lack of fit, LOF [33], were also derived to evaluate the best model. The model that produces the minimum values of AIC and LOF and the highest value of FIT is considered potentially the most useful and the best.

Fujita-Ban analysis

The Fujita-Ban analysis [34], based on an additivity principle, is a non-parametric approach and requires, relatively, a larger data-set. Additionally, the approach also needs certain substituent to occur two or more times at a given varying position in a molecule. Therefore, all compounds of Table 1 were considered together to enlarge the training-set in this approach. This may in turn give a better insight into the substitutional requirements for those analogues which are yet to be synthesized.

Partial Least Squares analysis

The partial least squares (PLS) [35-37] linear regression is a method suitable for overcoming the problems in MLR related to multicollinear or over-abundant descriptors. This is a modeling technique where information in the descriptor matrix X is projected onto a small number of latent variables (LV) called PLS components, which are linear combination of the original variables. The matrix Y is simultaneously used in estimating the "latent" variables in X that will be most relevant to predict the Y variables. All descriptor variables are preprocessed by autoscaling, using weights based on the variables' standard deviation and the data are mean-centered prior to PLS processing. Scaling of descriptors is necessary because the values have different orders of magnitude. Cross-validation was employed to select the used optimum number of LVs. With cross-validation, some samples were kept out of the calibration and used for prediction. The process was repeated so that each of the samples was kept out once. The predicted values of left-out samples were then compared to the observed values using predicted residual sum of squares (PRESS). The PRESS obtained in the cross-validation was calculated each time that a new LV was added to the model.

Applicability domain

The advantage of a QSAR model is based on its correct prediction ability for new compounds. A model is valid only within its training domain and new compounds must be assessed as belonging to the domain before the model is applied. The applicability domain (AD) is assessed by the leverage values for each compound [38,39]. The Williams plot (the plot of standardized residuals versus leverage values, h) can then be used for an instant and simple graphical detection of both the response outliers (Y-outliers) and structurally influential compound (X-outliers) in the model. In this plot, the AD is established inside a squared area within $\pm \beta\times$(s.d.) and a leverage threshold h^*. The threshold h^* is normally fixed at $3(k + 1)/n$ (n is the number of compounds in the analysis and k is the number of independent descriptors of the model) whereas $\beta = 2$ to 3. Prediction must be considered doubtful for compounds with a high leverage value ($h > h^*$). On the other hand, when the leverage value of a compound is lower than the threshold value, the probability of agreement between predicted and observed values is as high as that for the training-set compounds.

Results and Discussion

The Fujita-Ban (F.B.) analysis has been performed initially for the compounds in Table 1 to identify important structural features which make positive contribution to activity relative to the parent moiety so as to predict some new potential analogue of the series. A matrix comprising of 72 rows (compounds) and 20 columns (substituents at varying positions, including contribution due to parent moiety) was constructed for this purpose. The matrix element 1 or 0 represents, respectively, the presence or the absence of a substituent at given varying position of a compound. Similarly the value 1, considered for each compound, in the first column of this matrix reflects upon the contribution of parent moiety (a minimum substituted compound, S. No. **55** in Table 1). The documentation of this matrix, whose rows and columns represent, respectively, the data-points and independent variables, has been avoided here for the sake of brevity. Considering the pIC_{50} values as the dependent variable, the MRA revealed contributions of different substituents and parent moiety which are listed in Table 3 and highly significant statistical parameters as: n = 72, r = 0.999, s = 0.055 and F(20,51) = 1419.737.

Table.3 Fujita-Ban contribution of substituents and parent moiety to MMP-9 inhibition activity of caffeic acid amides

Position	Substituent	Contribution	Position	Substituent	Contribution
R_1	OH	0.076 (±0.02)	4-R_3	Cl	-0.087(±0.07)
R_2	OH	0.052 (±0.02)		F	-0.040 (±0.07)
2-R_3	Cl	-0.105 (±0.07)		Me	0.243 (±0.07)

	F	-0.060 (±0.07)		OH	0.300 (±0.07)
	Me	0.083 (±0.07)		OMe	0.260 (±0.07)
	OH	0.105 (±0.07)	n	1	-0.005 (±0.07)
	OMe	0.095 (±0.07)		2	-0.005 (±0.07)
3-R_3	Cl	-2.392 (±0.07)			
	F	-2.377 (±0.07)			
	Me	-2.370 (±0.07)			
	OH	-2.335 (±0.07)			
	OMe	-2.357 (±0.07)			
	Parent moiety, μ	8.056 (±0.05)			

In Table 3, the ±data within the parentheses, associated to substituent contributions obtained, are the 95% confidence intervals. The r^2-value accounts for 99.80% of the variance and the F-value is significant at 99% level [$F_{20,51}(0.01) = 2.250$]. The calculated pIC_{50} values, listed in Table 1, remained in close agreement with the observed ones. The plot showing the contributions of different substituents, relative to parent compound, is given in Figure 2.

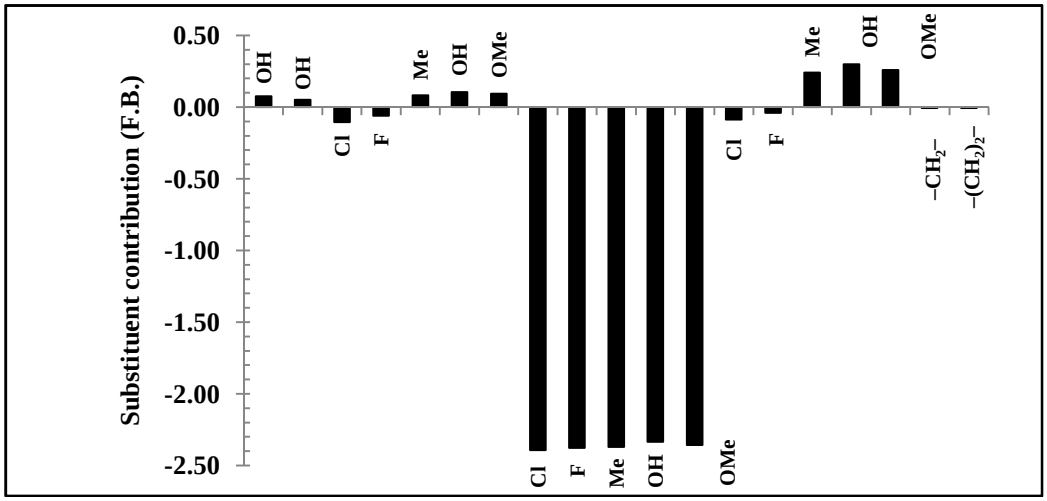

Figure.2 Plot of substituent contribution to MMP-9 inhibition activity, obtained from Fujita-Ban analysis.

From Table 3, the substituents that have a higher positive contribution to activity relative to substituents of the parent compound at different positions may easily be obtained. The –OH substitutions at R_1- and R_2-positions have shown positive contributions. Similarly, the electron-donor substituents at 2-R_3 and 4-R_3 revealed positive contributions in the order OH > OMe > Me, while both F and Cl at these positions have negative contributions. All substituents at 3-R_3 have shown negative contributions relative to H. Likewise, n = 1 and n = 2 indicating methylene insertion(s) between phenyl and amino group in the extended P1′ region also revealed negative contributions relative to n = 0 (no insertion). Thus, the appropriate substituents of varying positions which have highest positive contributions to the parent moiety may be selected for the future design of more active analogues of the series. The optimal MMP-9 inhibition activity seems to be manifested by the compounds in which $R_1 = R_2$ = OH, 2-R_3 = 4-R_4 = OH, OMe and Me and n = 0.

It is important to note that the Fujita-Ban approach, being non-parametric in nature, cannot extrapolate beyond the substituents of the original data-set while the parametric approach, attempted next, can do so. In the parametric approach, the data-set was divided into training-set and test-set which contain 54 and 18 compounds respectively. The test-set compounds were chosen in such way that they are representative of highly active, moderately active and poorly active congeners of the series. For this, the deviations of pIC_{50} values about their mean were calculated and absolute values of such deviations were arranged in ascending order for all the compounds. Then every fourth compound was selected to include in the test-set. In this way, the test-set has 18 out of 72 compounds and displayed sufficient activity variations and also represented structural diversity of the compounds. In the present case, compounds **1**, **3**, **4**, **8**, **9**, **12**, **20**, **22**, **31**, **32**, **35**, **36**, **37**, **44**, **46**, **49**, **59** and **67** (Table 1) are the representative constituents of the test-set. A total number of 428 2D-descriptors, computed from DRAGON software were subjected to elimination process. Those descriptors which were inter-correlated beyond 0.90 and showing a correlation of less than 0.1 with the pIC_{50}s were excluded at this stage. The left-over 70 significant descriptors were

scaled [40] further so that their values remain between 0 and 1. In this way, they would show equal influence in the QSAR models and none dominate each other like in the case of pre-scaled descriptors with larger or smaller values. The scaled descriptors were explored to develop models, through CP-MLR, which may explain MMP-9 inhibition actions of the compounds. A large number of models were generated in one-, two- and three-descriptors but only 7 models, each in three-descriptors, remained statistical significant which could explain up to 87.24% of variance in observed pIC_{50} values. The 10 participated descriptors in these models along with their brief description, average regression coefficients and the total incidence are given in Table 4.

Table.4 Identified descriptors[a], their physical meaning, average regression coefficient and incidence[b] in modeling of MMP-9 inhibition activity

S. No.	Name	Class	Physical meaning	Avg. reg. coefficient (incidence)
1	Jhetp	TOPO	Balaban-type index from polarizability weighted distance matrix.	-2.152 (4)
2	X4sol	TOPO	Solvation connectivity index chi-4.	0.581 (1)
3	PW4	TOPO	Path/walk 4 – Randic shape index.	1.159 (1)
4	BELe7	BCUT	Lowest eigenvalue no. 7 of Burden matrix/ weighted by atomic Sanderson electronegativities.	1.154 (1)
5	MATS4m	2DAUTO	Moran autocorrelation-lag 4/ weighted by atomic masses.	-0.887 (1)
6	MATS6e	2DAUTO	Moran autocorrelation-lag 6/ weighted by atomic Sanderson electronegativities.	-4.240 (7)
7	GATS1v	2DAUTO	Geary autocorrelation-lag 1/ weighted by atomic van der Waals volumes.	1.518 (1)
8	GATS3p	2DAUTO	Geary autocorrelation-lag 3/ weighted by atomic polarizabilities.	-0.746 (1)
9	C-001	ACF	Corresponds to CH3R/CH4.	-1.001 (3)
10	H-047	ACF	Corresponds to H attached to C1(sp^3)/ C0(sp^2)	1.155 (1)

[a]The descriptors have been identified from the models, emerged from CP-MLR protocol with a training-set of 54 compounds for MMP-9 inhibition activity; [b]The average regression coefficient of the descriptor corresponding to all models and the total number of its incidence. The arithmetic sign of the coefficient represents the actual sign of the regression coefficient in the models.

The developed models, in the increasing level of significance, are documented through Equations (1)-(7)

pIC_{50} = 8.754 –4.178(±0.275)MATS6e –0.703(±0.1764)C-001 +1.155(±0.210)H-047
n = 54, r = 0.911, s = 0.4694, F(3,50) = 81.111, Q^2_{LOO} = 0.797, Q^2_{L5O} = 0.782
r^2_{Test} = 0.774, r^2_{randY}(sd) = 0.225(0.076), FIT = 3.862, AIC = 0.256, LOF = 0.258 (1)

pIC_{50} = 10.310 –2.259(±0.279)Jhetp +0.581(±0.273)X4sol –4.324(±0.274)MATS6e
n = 54, r = 0.917, s = 0.454, F(3,50) = 87.633, Q^2_{LOO} = 0.802, Q^2_{L5O} = 0.799
r^2_{Test} = 0.805, r^2_{randY}(sd) = 0.217(0.098), FIT = 4.173, AIC = 0.240, LOF = 0.242 (2)

pIC_{50} = 8.699 +1.154(±0.185)BELe7 –3.931(±0.258)MATS6e –1.083(±0.164)C-001
n = 54, r = 0.920, s = 0.446, F(3,50) = 91.601, Q^2_{LOO} = 0.821, Q^2_{L5O} = 0.821
r^2_{Test} = 0.870, r^2_{randY}(sd) = 0.225(0.094), FIT = 4.362, AIC = 0.231, LOF = 0.233 (3)

pIC_{50} = 10.724 –1.944(±0.261)Jhetp –4.269(±0.260)MATS6e –0.746(±0.263)GATS3p
n = 54, r = 0.922, s = 0.440, F(3,50) = 94.368, Q^2_{LOO} = 0.827, Q^2_{L5O} = 0.829
r^2_{Test} = 0.837, r^2_{randY}(sd) = 0.215(0.095), FIT = 4.494, AIC = 0.225, LOF = 0.227 (4)

pIC_{50} = 8.600 –4.356(±0.253)MATS6e +1.517(±0.219)GATS1v –1.218(±0.160)C-001
n = 54, r = 0.928, s = 0.424, F(3,50) = 102.978, Q^2_{LOO} = 0.833, Q^2_{L5O} = 0.833
r^2_{Test} = 0.765, r^2_{randY}(sd) = 0.216(0.097), FIT = 4.904, AIC = 0.209, LOF = 0.211 (5)

pIC_{50} = 10.918 –1.942(±0.241)Jhetp –0.887(±0.214)MATS4m –4.061(±0.242)MATS6e
n = 54, r = 0.933, s = 0.409, F(3,50) = 111.846, Q^2_{LOO} = 0.846, Q^2_{L5O} = 0.856
r^2_{Test} = 0.786, r^2_{randY}(sd) = 0.234(0.102), FIT = 5.326, AIC = 0.194, LOF = 0.196 (6)

pIC_{50} = 10.379 –2.465(±0.255)Jhetp +1.159(±0.274)PW4 –4.563(±0.254)MATS6e
n = 54, r = 0.934, s = 0.407, F(3,50) = 113.306, Q^2_{LOO} = 0.848, Q^2_{L5O} = 0.855
r^2_{Test} = 0.808, r^2_{randY}(sd) = 0.212(0.088), FIT = 5.396, AIC = 0.192, LOF = 0.194 (7)

In above equations, the F-values remained significant at 99% level [$F_{3,50}$ (0.01) = 4.199] and the standard errors of regression coefficients (± data within the parentheses) were significant at more than 95% level. The indices Q^2_{LOO} and Q^2_{L5O} (> 0.5) have accounted for internal robustness of the developed models while the index r^2_{Test} greater than 0.5 specified that the selected test-set is accountable for external validation of these models. The low r^2_{randY}(sd) value revealed that none of the 100 activity randomized models is superior to that of original model.

In Equations (1)-(7), the descriptors, X4sol, PW4 and Jhetp (from TOPO class) stand, respectively, for solvation connectivity index chi-4, Randic shape index (path/walk 4) and Balaban-type index from polarizability weighted distance matrix. The descriptors BELe7 (from BCUT class) represents the lowest eigenvalue no. 7 of Burden matrix/ weighted by atomic Sanderson electronegativities. The descriptor is derived for a hydrogen-included molecular graph and calculated from Burden matrix whose diagonal elements are the atomic Sanderson electronegativities; the off-diagonal elements corresponding to pair of bonded atoms are the square roots of conventional bond order and all other matrix elements are set at 0.001. The descriptors, MATS*kw* and GATS*kw* (from 2DAUTO class) are the Moran and Geary spatial autocorrelations respectively. These descriptors are calculated on a hydrogen-depleted molecular graph, in which the weighting component, *w* is the atomic property such as atomic masses (m), or atomic van der Waals volumes (v) or atomic Sanderson electronegativities (e) or atomic polarizabilities (p), and *k* is the lag (path). Finally, the descriptors, C-001and H-047 (from ACF class) correspond, respectively, to the functionalities CH3R/CH4 and H attached to C1(sp^3)/ C0(sp^2). The signs of the regression coefficients have indicated the direction of influence of explanatory variables in a given model; the positive regression coefficient associated to a descriptor will augment the inhibition activity of a compound while the negative coefficient will cause detrimental effect to it. Equations (6) and (7), being the most significant amongst all emerged models, were retained for further discussion. The calculated MMP-9 inhibition activities, using these equations, have been listed in Table 1. The same remained in parity with the observed ones. The descriptors, Jhetp and MATS6e, participated in both these equations, have shown their negative influence on MMP-9 inhibition activity. Thus the lower polarizability and electron deficiency associated with lag (path)-6 are conducive in improving the inhibition action of a compound. Additionally, the lower value of atomic mass weighted lag-4 of Moran autocorrelation (MATS4m; Equation 6) or the higher ratio of the atomic path count over the atomic walk count of the order 4 (PW4; Equation 7) is also desirable.

Further, the PLS analysis was carried out on the 10 descriptors, identified through CP-MLR, to facilitate the development of a 'single window' structure-activity model and to identify the potentiality of these descriptors in explaining the MMP-9 inhibition actions of caffeic acid amides. The analysis also provides an opportunity to make a comparison of relative significance among these descriptors. The fraction contributions that were obtained from the normalized regression coefficients of the descriptors allow such comparison within the modeled activity. For PLS analysis, the descriptors have been autoscaled (zero mean and unit s.d.) to give each one of them equal weight in the analysis. In the PLS cross-validation, three components remained optimum for 10 descriptors and explained 87.05% of variance in the observed pIC_{50}s. The PLS equation, its regression statistics, the MLR-like PLS coefficients of 10 descriptors and their fractional contribution to activity are given in Table 5.

Table.5 PLS and MLR-like PLS equation from the descriptors of CP-MLR identified models for MMP-9 inhibition activity

A: PLS Equation

PLS components	PLS coefficient (s. e.)[a]
Component-1	0.697 (0.042)
Component-2	-0.252 (0.035)
Component-3	-0.089 (0.040)
Constant	7.505

B: PLS regression statistics

Symbol	Estimate
n	54
r	0.933
s	0.408
F	112.578
Q^2_{LOO}	0.848
Q^2_{L5O}	0.840
r^2_{Test}	0.831

C: MLR-Like PLS Equation

S. No.	Descriptor	MLR-like coefficient[b]	F. C. (order)[b]	S. No.	Descriptor	MLR-like coefficient[b]	F.C. (order)[b]
1	Jhetp	-0.482	-0.058 (5)	7	GATS1v	-0.093	-0.013 (10)

2	X4sol	-0.278	-0.033 (8)	8	GATS3p	-0.483	-0.056 (6)
3	PW4	-0.425	-0.047 (7)	9	C-001	-0.492	-0.094 (4)
4	BELe7	0.577	0.094 (3)	10	H-047	0.120	0.019 (9)
5	MATS4m	-1.186	-0.154 (2)		Constant	10.394	
6	MATS6e	-3.646	-0.434 (1)				

[a]Regression coefficient of PLS factor and its standard error; [b]Coefficients of MLR-like PLS equation in terms of descriptors for their original values; F.C. is fraction contribution of regression coefficient, computed from the normalized regression coefficients obtained from the autoscaled (zero mean and unit standard deviation) data.

The calculated activity values, listed in Table 1, are found in close agreement with the observed ones. Figure 3 shows a plot of the fraction contribution of normalized regression coefficients of these descriptors to the activity.

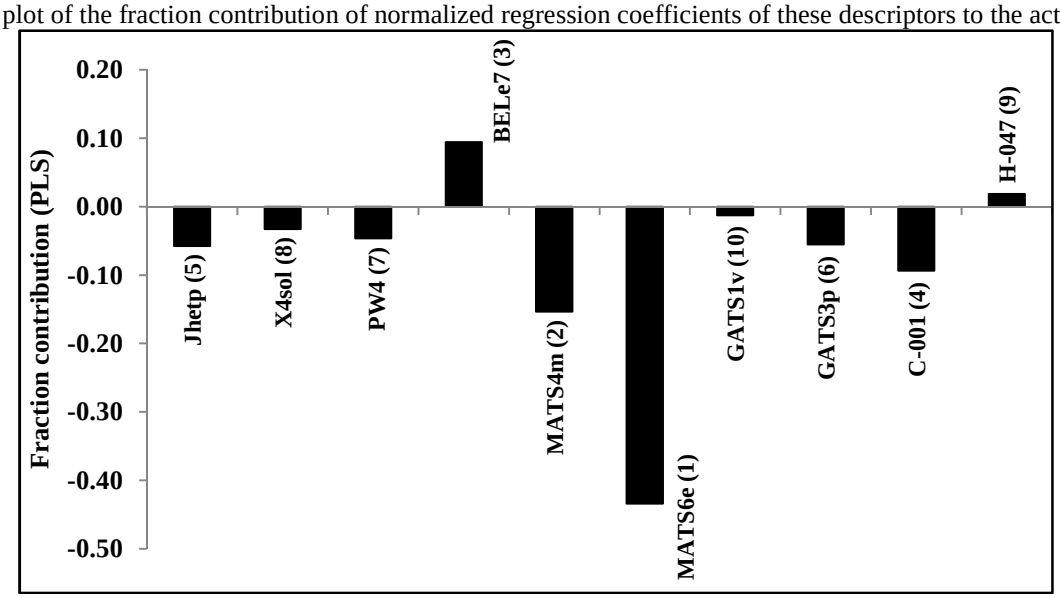

Figure.3 Plot between fraction contribution of MLR-like PLS coefficients (normalized) and 10 identified descriptors (Table 5) associated with inhibition actions of caffeic acid amide derivatives.

Different orders, indicating the level of significance of participated descriptors, are also included in Table 5. For a given descriptor, lower is the order higher would be the significance in addressing the biological activity. The descriptors having positive contribution will augment the activity and their higher values are desirable to further improve it. On the other hand, the descriptors having negative contribution will diminish the activity. The lower or more negative values of such descriptors may, therefore, enhance the activity of a compound. The analysis also suggested MATS6e as the most influential descriptor for modeling the activity of the compounds (descriptor S. No. 6 in Table 5). The remaining descriptors, in decreasing order of significance, are MATS4m, BELe7, C-001, Jhetp, GATS3p, PW4, X4sol, H-047 and GATS1v (descriptors S. No. 5, 4, 9, 1, 8, 3, 2, 10 and 7 in Table 5). The investigation of new potential compounds prior to their synthesis is one of the important facets of a QSAR study. It will minimize the time and cost coupled with identifying new leads. In the present case, the rationalization of new analogues is based on the inferences drawn from the Fujita-Ban approach, highest significant Equations (6)-(7) and PLS analysis. For this, a virtual screening was performed by insertion, deletion and substitution of different substituents on the original compounds. The effects of the structural modifications, reflected mainly through the most significant descriptors on the biological activity were investigated. A few congeners (Figure 4) having much higher activity profiles, compared to the highest potent compounds reported in the original series (Table 1), are suggested for further exploration. These are listed in Table 6 along with their modeled pIC_{50} values.

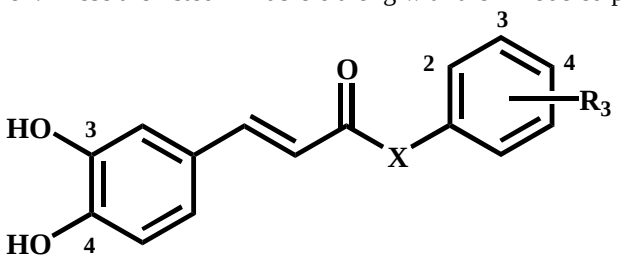

Figure.4 Predicted caffeic acid amide derivatives.

Table.6 Predicted compounds and their modeled pIC_{50} values

S. No.	X	R_3	Predicted pIC_{50} (M)			
			F.B.	Eq. (6)	Eq. (7)	PLS
73	-NH-	2-OH, 4-OMe	8.55	9.15	9.15	9.09
74	-NH-	2,4-diOMe	8.54	9.25	9.23	9.14
75	-O-	2-OH, 4-OMe	--	10.88	11.28	9.63
76	-O-	2,4-diOMe	--	11.12	11.47	9.87

The applicability domain (AD) was analyzed for the models resulted from all 72 compounds of the series. It is illustrated through the Williams plot, obtained between standardized residuals and leverage (h_i) values. For this purpose, the descriptors of the most significant Equations (6) and (7) were used to derive corresponding models based on whole data-set. The resulted models are given through Equations (8) and (9) while standardized residuals and leverage values, calculated in conjunction with them, are used to ascertain their ADs.

pIC_{50} = 10.699 –1.803(±0.231)Jhetp –0.650(±0.210)MATS4m –4.109(±0.220)MATS6e

n = 72, r = 0.924, s = 0.431, F(3,68) = 132.411, Q^2_{LOO} = 0.834, Q^2_{L5O} = 0.832

r^2_{randY}(sd) = 0.193(0.072), FIT = 4.904, AIC = 0.207, LOF = 0.208 (8)

pIC_{50} = 10.330 –2.259(±0.241)Jhetp +0.932(±0.268)PW4 –4.508(±0.229)MATS6e

n = 72, r = 0.927, s = 0.424, F(3,68) = 137.464, Q^2_{LOO} = 0.839, Q^2_{L5O} = 0.842

r^2_{randY}(sd) = 0.180(0.075), FIT = 5.091, AIC = 0.201, LOF = 0.202 (9)

The limits of normal values for the standardized residuals (response or Y-outliers) were set as ±3×s.d. while leverage threshold as h*. The graphical representations for these models are given in Figure 5.

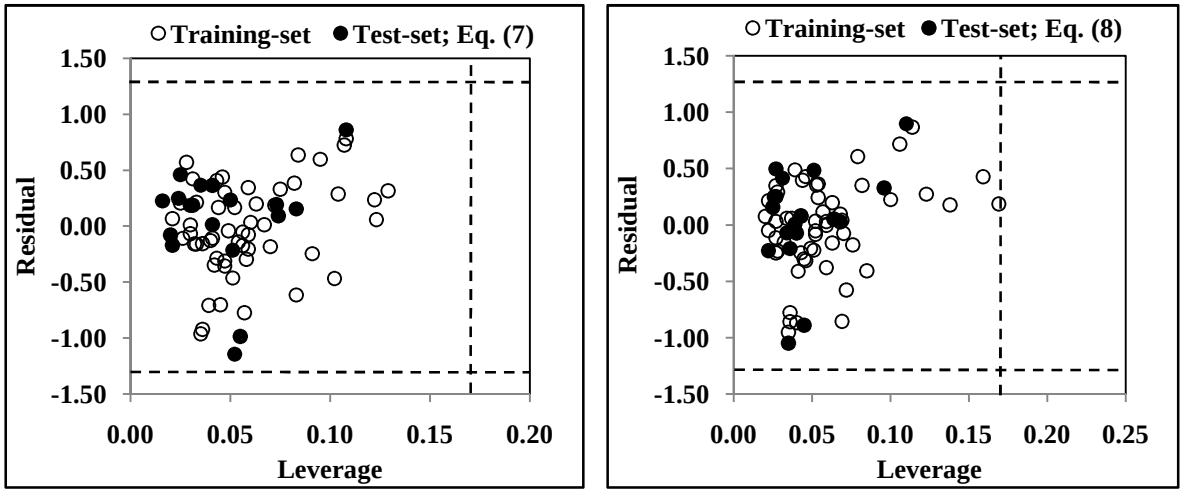

Figure.5 Williams plots corresponding to Equations (7) and (8) for whole data-set for inhibition actions of caffeic acid amides listed in Table 1 (h* = 0.167 and residual limits are ±3.0×s.d.).

For both the training- and test-set compounds, the suggested models match the high quality parameters with good fitting power and the capability of assessing external data. Moreover, all data-points were present within the AD of both these models. This implies that the models under consideration are able to evaluate, both the training-set and test-set compounds, correctly.

Conclusion

The MMP-9 inhibition activity of caffeic acid amides is quantitatively analyzed through non-parametric and parametric approaches. The non-parametric approach involves Fujita-Ban analysis provided the contributions of different substituents relative to parent compound. The appropriate substituent, which showed highest positive contribution to inhibition activity were selected to design more active analogues. In the present case, the optimal MMP-9 inhibition activity appears to be exhibited by compounds in which $R_1 = R_2$ = OH, 2-R_3 = 4-R_4 = OH, OMe and Me and n = 0 (Figure 1). The parametric approach utilized molecular 2D-descriptors to develop statistically

validated models which were able to explain inhibition actions of the compounds. 7 Models, each in three descriptors, were filtered out from the protocol based approach, CP-MLR. They shared a total number of 10 2D-descriptors but two of these models, exhibiting highest levels of significance, were discussed further. The descriptors involved in them were Jhetp, MATS6e, MATS4m and PW4 which have significantly addressed the MMP-9 inhibition activity of the compounds. The Jhetp and MATS6e imparted their negative influence on activity. Thus the lower polarizability and electron deficiency associated with lag (path)-6 are conducive in improving the inhibition action of a compound. Also, the lower bulk associated to lag-4 (MATS4m) or the higher value of PW4 is also desirable. The PLS analysis revealed a 'single window' structure-activity model from 10 identified descriptors and helped to make a comparison of relative significance among these descriptors. The developed model involving three optimum components could explain 87.05% of variance in the observed activity values of the compounds. The strategies delineated by the Fujita-Ban approach, statistically validated models and PLS analysis, facilitated in exploring new potential analogues for further exploration. Applicability domain analysis revealed that the suggested models have acceptable predictability. All the compounds were within the applicability domain of the proposed models and were evaluated correctly.

Acknowledgements

Financial support provided by the University Grants Commission, New Delhi is thankfully acknowledged. The Author is grateful to his host Institution for providing necessary facilities to complete this work and conveys his deep appreciation to Dr. Y.S. Prabhakar, Scientist E II, Medicinal Process Chemistry Division, CDRI, Lucknow for his continuous support.

References

1. H Nagase; JF Woessner Jr *J Biol Chem* **1999**, *274*, 21491-21494.
2. JR MacDougall; LM Matrisian *Cancer Metastasis Rev* **1995,** *14*, 351-362.
3. WG Stetler-Stevenson *J Clin Invest* **1999,** *103*, 1237-1241.
4. RT Aimes; JP Quigley *J Biol Chem* **1995**, *270*, 5872-5876.
5. T Crabbe; JP O'Connell; BJ Smith; AJP Docherty *Biochemistry* **1994**, *33*, 14419-14425.
6. B Lovejoy; A Cleasby; AM Hassell; K Longley; MA Luther; D Weigl; G McGeehan; AB McElroy; D Drewry; MH Lanbert; SR Jordan *Science* **1994**, *263*, 375-377.
7. PR Gooley; JF O'Connell; AI Marcy; GC Cuca; SP Salowe; BL Bush; JD Hermes; CK Esser; WK Hagmann; JP Springer; BA Johnson *Nat Struct Biol* **1994**, *1*, 111-118.
8. NG Li; ZH Shi; YP Tang; JA Duan *Curr Med Chem* **2009**, *16*, 3805-3827.
9. WH Park; SH Kim; CH Kim *Toxicology* **2005**, *207*, 383-390.
10. D Grunberger; R Banerjee; K Eisinger; EM Oltz; L Efros; M Caldwell; V Estevez; K Nakanishi *Experientia* **1988**, *44*, 230-232.
11. T Nagaoka; AH Banskota; Y Tezuka; I Saiki; S Kadota *Bioorg Med Chem* **2002**, *10*, 3351-3359.
12. TW Chung; SK Moon; YC Chang; JH Ko; YC Lee; G Cho; SH Kim; JG Kim; CH Kim *FASEB J* **2004**, *18*, 1670-1681.
13. T Nakawaza; K Ohsawa *J Nat Prod* **1998**, *61*, 993-996.
14. EU Graefe; M Veit *Phytomedicine* **1999**, *6*, 239-246.
15. N Celli; B Mariani; LK Dragani; S Murzilli; C Rossi; DJ Rotilio ncbi *Chromatogr B* **2004**, *810*, 129-136.
17. P Rajan; I Vedernikova; P Cos; DV Berghe; K Augustyns; A Haemers *Bioorg Med Chem Lett* **2001**, *11*, 215-217.
18. ZH Shi; NG Li; QP Shi; H Tang; YP Tang; W Li; L Yin; JP Yang; JA Duan *Bioorg Med Chem Lett* **2013**, *23*, 1206-1211.
19. Chemdraw ultra 6.0 and Chem3D ultra, Cambridge Soft Corporation, Cambridge, USA. http://www.cambridgesoft.com
20. DRAGON software (version 3.0-2003) by R Todeschini; V Consonni; A Mauri; M Pavan, Milano, Italy. http://www.talete.mi.it/dragon.htm.
21. YS Prabhakar *QSAR Comb Sci* **2003**, *22*, 583-595.
22. S Sharma; YS Prabhakar; P Singh; BK Sharma *Eur J Med Chem* **2008**, *43*, 2354-2360.
23. S Sharma; BK Sharma; SK Sharma; P Singh; YS Prabhakar *Eur J Med Chem* **2009**, *44*, 1377-1382.
24. BK Sharma; P Pilania; P Singh; YS Prabhakar *SAR QSAR Environ Res* **2010**, *21*, 169-185.
25. BK Sharma; P Singh; K Sarbhai; YS Prabhakar *SAR QSAR Environ Res* **2010**, *21*, 369-388.
26. BK Sharma; P Pilania; K Sarbhai; P Singh; YS Prabhakar *Mol Divers* **2010**, *14*, 371-384.
27. BK Sharma; P Singh; M Shekhawat; K Sarbhai; YS Prabhakar *SAR QSAR Environ Res* **2011**, *22*, 365-383.
28. S-S So; M Karplus *J Med Chem* **1997**, *40*, 4347-4359.
29. YS Prabhakar; VR Solomon; RK Rawal; MK Gupta; SB Katti *QSAR Comb Sci* **2004**, *23*, 234-244.
30. H Akaike. In *Second international symposium on information theory* eds. BN Petrov; F Csaki, Akademiai Kiado; Budapest **1973**; pp. 267-281.

31. H Akaike *IEEE Trans Autom Control* **1974**, *AC-19*, 716-723.
32. H Kubinyi *Quant Struct-Act Relat* **1994**, *13*, 285-294.
33. H Kubinyi H. *Quant Struct-Act Relat* **1994**, *13*, 393-401.
34. J Friedman *Technical report no 102 Laboratory for computational statistics*, Stanford University, November **1988**.
35. T Fujita; T Ban *J Med Chem* **1971**, *14***,** 148-152.
36. S Wold *Technomet* **1978**, *20*, 397-405.
37. N Kettaneh; A Berglund; S Wold *Comput Stat Data Anal* **2005**, *48*, 69-85.
38. L Stahle; S Wold. In *Progress in Medicinal Chemistry*, eds. GP Ellis; GB West, Elsevier Science Publishers BV, Amsterdam, **1988**; pp. 291-338.
39. P Gramatica *QSAR Comb Sci* **2007**, *26*, 694-701.
40. L Eriksson; J Jaworska; AP Worth; MTD Cronin; RM McDowell; P Gramatica *Environ Health Persp* **2003**, *111*, 1361-1375.
41. A Golbraikh; A Tropsha *J Mol Graph Model* **2002**, *20*, 269-276.