Digest Journal of Nanomaterials and Biostructures Vol. 3, No. 3, September 2008, p. 147- 158

THREE-DIMENSIONAL QSAR STUDY OF 2,4 - DISUBSTITUTED-PHENOXY ACETIC ACID DERIVATIVES AS A $CRTh_2$ RECEPTOR ANTAGONIST: USING THE k-NEAREST NEIGHBOR METHOD

Abhishek K. Jain[a], Nimita Manocha[c] , V. Ravichandran[a] V. K. Mourya[b], R. K. Agrawal[a*]

[a] *Pharmaceutical Chemistry Research Laboratory, Dept. of Pharmaceutical Sciences,*
Dr. Hari Singh Gour University, Sagar (M.P) - 470 003, India.
[b] *Govt. College of Pharmacy, Osmanpura, Aurangabad, Maharashtra, India.*
[c]*Swami Vivekanand College of Pharmacy, Indore (M.P.)*

In pursuit of better $CRTh_2$ receptor antagonist agents, 3D- QSAR studies were performed on a series of 2,4 -disubstituted-phenoxy acetic acid derivatives. In this paper we report a novel three-dimensional QSAR approach, kNN-MFA, developed based on principles of the k-nearest neighbor method combined with various variable selection procedures. The kNN-MFA approach was used to generate models by all three different methods and predict the activity of test molecules through each of these models. The q^2, pred_r^2, Vn and k value of kNN-MFA with SW, SA & GA were (0.8392, 0.7059, 2/2) (0.6725, 0.6716, 2/4) and (0.6832, 0.6716, 2/4) although there are no common descriptors among these three methods, SW kNN-MFA method have better q2 (0.8392) and pred_r2 (0.7059) than other two methods, model validation correctly predicts activity 83.9% and 70.5% for the training and test set respectively. It uses 2 steric descriptors with 2 k nearest neighbor to evaluate activity of new molecule, So model generated by SA kNN-MFA are best model.

(Received July 20, 2008; accepted July 25, 2008)

Keywords: QSAR; kNN-MFA, 2,4 -disubstituted-phenoxy acetic acid.

1. Introduction

The lipid mediator prostaglandin D2 (PGD_2) is implicated in various inflammatory disease including asthma.[1-2]. PGD_2 exhibit its biological responses by activating two seven transmembrane (7TM) G-protein coupled receptors, the D-prostanoid receptor DP and the chemoattractant receptor homologous-molecule expressed on T-helper-type-2 cells ($CRTh_2$), which are linked to different signaling pathways. $CRTh_2$ negatively regulates adenylyl cyclases through Gi proteins, mobilizes intracellular calcium, and stimulates phosphoinositide 3-kinase, mitogen-activated protein kinases and phospholipase C [2].

Endogenous agonists for $CRTh_2$ include PGD_2 and a number of its metabolites, notably 13,14-dihydro-15-keto PGD_2 (DK-PGD_2), which is selective for $CRTh_2$ over DP_1. Activation of $CRTh_2$ promotes chemotaxis of Th_2 cells, eosinophils and basophils, as well as degranulation of eosinophils and cytokine release from Th_2 cells.[2]

The orally available small molecule ramatroban also antagonizes $CRTh_2$ with a potency sufficient to account at least in part for the beneficial clinical effects of ramatroban[3-5]. However, the inability of ramatroban to selectively inhibit one receptor potent and selective $CRTh_2$ antagonists would be desirable to explore the involvement of PGD_2 and $CRTh_2$ in allergic and atopic conditions.

*Corresponding author: dragrawal2001@yahoo.co.in

The present group of authors has developed a few quantitative structure-activity relationship models to predict biological activity of different group of compounds. [6-12]

Many different approaches to QSAR have been developed over the years. The rapid increase in three-dimensional structural information (3D) of bioorganic molecules, coupled with the development of fast methods for 3D structure alignment (e.g. active analogue approach), has led to the development of 3D structural descriptors and associated 3D QSAR methods. The most popular 3D QSAR methods are comparative molecular field analysis (CoMFA)[13] and comparative molecular similarity analysis (CoMSIA).[14] The CoMFA method involves generation of a common threedimensional lattice around a set of molecules and calculation of the steric and electrostatic interaction energies at the lattice points. The interaction energies are numerically very high when a lattice point is very close to an atom and special care needs to be taken in order to avoid problems arising because of this. The CoMSIA method avoids these problems by using similarity function represented as Gaussian. This information around the molecule is converted into numerical data using the partial least squares (PLS) method that reduces the dimensionality of data by generating components. However, a major disadvantage is that PLS attempts to fit a linear curve among all the points in the data set. Further, the PLS method does not offer scope for improvement in results. It has been observed from several reports that the predictive ability of PLS method is rather poor due to fitting of a linear curve between the available points. In the case of the CoMSIA method, molecular similarity is evaluated and used instead of molecular field, followed by PLS analysis.

Variable selection methods have also been adopted for optimal region selection in 3D QSAR methods and shown to provide improved QSAR models as compared to the original CoMFA technique. For example, GOLPE[15] was developed using chemometric principles, and q2-GRS was developed on the basis of independent analyses of small areas (or regions) of near-molecular space to address the issue of optimal region selection in CoMFA.[16] These considerations provide an impetus for the development of fast, generally nonlinear, variable selection methods for performing molecular field analysis. We report here the development of a new method (kNN-MFA) that adopts a k-nearest neighbor principle for generating relationships of molecular fields with the experimentally reported activity. This method utilizes the active analogue principle that lies at the foundation of medicinal chemistry.

2. Methodology

We hereby report the models, as generated by kNN-MFA in conjunction with simulated annealing (SA), genetic algorithm (GA), and stepwise (SW) forward variable selection methods. In the kNN-MFA method, several models were generated for the selected members of training and test sets, and the corresponding best models are reported herein.

VLife Molecular Design Suite (VLifeMDS), allows user to choose probe, grid size, and grid interval for the generation of descriptors. The variable selection methods along with the corresponding parameters are allowed to be chosen, and optimum models are generated by maximizing q^2. k-nearest neighbor molecular field analysis (kNN-MFA) requires suitable alignment of given set of molecules. This is followed by generation of a common rectangular grid around the molecules. The steric and electrostatic interaction energies are computed at the lattice points of the grid using a methyl probe of charge +1. These interaction energy values are considered for relationship generation and utilized as descriptors to decide nearness between molecules. The term descriptor is utilized in the following discussion to indicate field values at the lattice points. The optimal training and test sets were generated using the sphere exclusion algorithm.[17] This algorithm allows the construction of training sets covering descriptor space occupied by representative points. Once the training and test sets were generated, kNN methodology was applied to the descriptors generated over the grid.

Nearest Neighbor (kNN) Method

The kNN methodology relies on a simple distance learning approach whereby an unknown member is classified according to the majority of its k-nearest neighbors in the training set. The

nearness is measured by an appropriate distance metric (e.g., a molecular similarity measure calculated using field interactions of molecular structures). The standard kNN method is implemented simply as follows:[18]Calculate distances between an unknown object (u) and all the objects in the training set; select k objects from the training set most similar to object u, according to the calculated distances; and classify object u with the group to which the majority of the k objects belongs. An optimal k value is selected by optimization through the classification of a test set of samples or by leave-one out cross-validation.

kNN-MFA with Simulated Annealing

Simulated annealing (SA) is the simulation of a physical process, 'annealing', which involves heating the system to a high temperature and then gradually cooling it down to a preset temperature (e.g., room temperature). During this process, the system samples possible configurations distributed according to the Boltzmann distribution so that at equilibrium, low energy states are the most populated.

kNN-MFA with Stepwise (SW) Variable Selection

This method employs a stepwise variable selection procedure combined with kNN to optimize the number of nearest neighbors (k) and the selection of variables from the original pool as described in simulated annealing.

kNN-MFA with Genetic Algorithm

Genetic algorithms (GA) first described by Holland[19] mimic natural evolution and selection. In biological systems, genetic information that determines the individuality of an organism is stored in chromosomes. Chromosomes are replicated and passed onto the next generation with selection criteria depending on fitness.

Cross-Validation Using Weighted k-Nearest Neighbor.

(1) A molecule in the training set was eliminated, and its biological activity was predicted as the weighted average activity of the k most similar molecules (eq 1). The similarities were evaluated as the inverse of Euclidean distances between molecules (eq 2) using only the subset of descriptors corresponding to the current trial solution.

$$\mathrm{wi} = \frac{\mathrm{Exp}\ (\text{-dj})}{\sum \mathrm{Exp}\ (\text{-dj})}$$

k-Nearest neighbour

$$\hat{y}_i = \sum w_i y_i \qquad (1)$$

$$d_{ij} = \left(\sum_{k=1}^{vn} (X_{i,k} - X_{j,k})^2\right)^{1/2} \qquad (2)$$

(2) Step 1 was repeated until every molecule in the training set has been eliminated and its activity predicted once.
(3) The cross-validated r^2 (q^2) value was calculated using eq 3, where *yi* and *y*^*i* are the actual and predicted activities of the *i*th molecule, respectively, and *y*mean is the average

activity of all molecules in the training set. Both summations are over all molecules in the training set. Since the calculation of the pairwise molecular similarities, and hence the predictions, were based upon the current trial solution, the q^2 obtained is indicative of the predictive power of the current kNN-MFA model.

$$q^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - y_{mean})^2} \quad (3)$$

External Validation.

The predicted r^2 (pred_r^2) value was calculated using eq 4, where *yi* and *y^i* are the actual and predicted activities of the i^{th} molecule in test set, respectively, and y_{mean} is the average activity of all molecules in the training set. Both summations are over all molecules in the test set. The pred_r^2 value is indicative of the predictive power of the
current kNN-MFA model for external test set.

$$pred\text{-}r^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - y_{mean})^2} \quad (4)$$

Randomization Test.

To evaluate the statistical significance of the QSAR model for an actual data set, we have employed a one-tail hypothesis testing.[20-21] The robustness of the QSAR models for experimental training sets was examined by comparing these models to those derived for random data sets. Random sets were generated by rearranging biological activities of the training set molecules. The significance of the models hence obtained was derived based on calculated Zscore.[20-21]

Evaluation of the QSAR Models.

The QSAR models were evaluated using following statistical measures: *n*, number of observations (molecules); Vn, number of descriptors; *k*, number of nearest neighbors; q^2, cross-validated r^2 (by the leave-one-out method); pred_r^2, predicted r^2 for the external test set; Zs_{core}, the Z score calculated by q^2 in the randomization test; best_ran_q^2, the highest q^2 value in the randomization test; and R, the statistical significance parameter obtained by the randomization test.

Experimental

The biological activities of all 31 compounds were collected from the reported series[5] (Table 3) . All the thirty one compounds were built on workspace of molecular modeling software VLifeMDS, which is a product VLife Sciences Pvt Ltd.,India[22]. We here by report the models, as generated by kNN-MFA in conjunction with simulated annealing (SA), genetic algorithm (GA), and stepwise (SW) forward variable selection methods (Table 4) .In the kNN-MFA method, several models were generated for the given or selected members of training and test sets, and the corresponding best models are reported herein. The method described above has been implemented in a software, VLife Molecular Design Suite (VLifeMDS),[22] which allows user to choose probe, grid size, and grid interval for the generation of descriptors. The variable selection methods along with the corresponding parameters are allowed to be chosen, and optimum models are generated by maximizing *q*2.

Steps involved in kNN-MFA method

1. Molecules are optimized before alignment optimization is done by MOPAC energy minimization and optimization is necessary process for proper alignment of molecules around template.

2. kNN-MFA method requires suitable alignment of given set of molecules, alignment are template based

3. This is followed by generation of common rectangular grid around the molecules, the steric and electrostatic interaction energies are computed at the lattice points of the grid using a methyl probe of charge +1.

4**.** The optimal training and test set were generated using sphere exclusion method.

5. Model was generated by various kNN methods, and models validated internally and externally by leave one out, external validation (q^2, r^2).

6. Predict the activity of test set of compounds.

Training and test set

Selection of training and test set by using sphere exclusion method for choosing uniformly distributed molecules in both sets. (compound number 9, 19, & 20 were selected for test set)

Table 1- Uni-Column statistics for training set

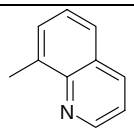
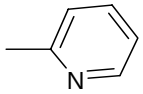
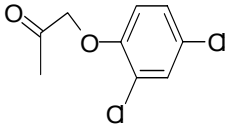
Col. name	Average	Max	Min	stdDev	sum
Column	4.0150	5.8539	1.8928	0.524	112.4188

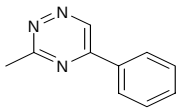
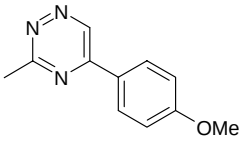
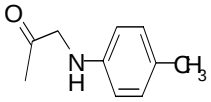
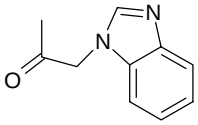
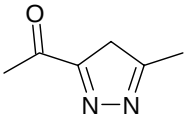
Table 2- Uni-Column statistics for test set

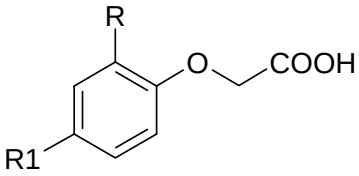
Col. name	Average	Max	Min	stdDev	sum
Column	4.2827	5.8239	2.7055	0.539	12.8482

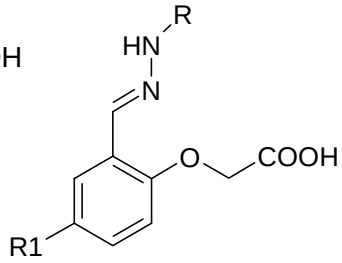
The max of the test should be less than or equal to max of train set and the min of the test should be greater than or equal to min of train set.

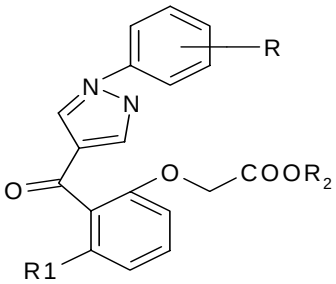
Table 3. Structure, $CRTh_2$ antagonist activity (IC_{50} in nm) of 2,4-disubstituted-phenoxy acetic acid derivatives.

Compound	**R1**	**R**	**IC_{50} (in nm) *binding*[a]**
1	Br		54
2	Br		255
3	H		555

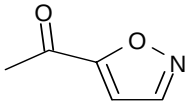
4	Br		89
5	H		1580
6	H		3660
7	H		12800
8	*H*		11200







Compound	R_1	R	IC_{50} (in nm) binding[a]
9	H	HCO-	1970
10	Br	HCO-	124
11	Br	CH3CO-	510
12	Br	PhCO-	53
13	Br		391



14	Br	t-Bu	61
15	Br	HOCH2-	886

Compound 9-15 Compound 15-31

Compound	R2	R1	R	IC_{50} (in nm) binding[a]
16	H	H	H	510
17	H	F	H	108
18	H	Cl	H	15
19	H	Br	H	1.5
20	H	CH_3	H	48
21	H	OCH_3	H	151
22	H	NO_2	H	47
23	H	Ph	H	42
24	H	4-Cl-C_6H_4	H	37
25	H	3,5-F-C_6H_3	H	13
26	Et	Br	*p*-OCH_3	165
27	H	Br	*p*-Cl	3.6
28	H	Br	*p*-Br	1.5
29	H	Br	*p*-Br	1.4
30	H	Br	*m*-Br	3.4
31	H	Br	*o*-Br	1.9

Table 4 - Comparison of the various kNN-MFA models for $CRTh_2$ receptor antagonist

kNN-MFA Method	Descriptors	Statistical Parameter
Stepwise (SW) Variable Selection	S_1580 (-0.0082 -0.0080) S_394 (-0.0023 -0.0023)	q^2=0.839, pred_r^2=0.7059, Z score= 3.195, best_rand_q^2=0.194, pred_r^2se=0.8642, α = 〉0.001, k/Vn=2/2, n= 28, test size= 3
Simulated Annealing (SA)	S_21 (-0.0010 -0.0006) S_183 (-0.0029 -0.0029) E_1244(-0.8367 -0.6487) E_254 (-0.0103 0.0017)	q^2=0.672, pred_r^2=0.6502, Z score= 4.185, best_rand_q^2= -0.1003, pred_r^2se=0.9425

		α = 〉0.001, k/Vn=2/4, n= 28, test size= 3
Genetic Algorithm (GA)	E_613 (-0.0124 0.0083) E_1481 (0.0121 0.0306) S_1104 (-0.0575 -0.0450) E_1415 (0.2379 0.4065)	q^2=0.6832, pred_r^2=0.6716, Z score= 4.986, best_rand_q^2= -0.203, pred_r^2se=0.9133 α = 0.001, k/Vn=2/4, n= 28, test size= 3

n, number of observations (molecules); Vn, number of descriptors; k, number of nearest neighbors; q2, cross-validated r2 (by the leave-one out method); pred_r2, predicted r2 for the external test set; Zscore, the Z score calculated by q2 in the randomization test; best_ran_q2, the highest q2 value in the randomization test; and R, the statistical significance parameter obtained by the randomization test.

3. Results and discussion

A data set of 31 compounds of reported series for $CRTh_2$ receptor antagonist activity was used for the present QSAR study[6] (Table 3). The QSAR studies of the 2,4-disubstituted-phenoxy acetic acid derivatives series resulted in several QSAR equations. The three best equations by all three kNN method are (with statistical parameter) given in table 4.

The calculated and predicted activities of the training and test set of compounds by models generated through all three methods are shown in table 5 and table 6.

Table 5 - Comparison of predicted activities of various kNN-MFA method for training set of compounds

Com.	Actual Activity*	Predicted Activity by kNN-MFA with SW	Predicted Activity by kNN-MFA with SA	Predicted Activity by kNN-MFA with GA
1	4.268	3.671	3.873	3.218
2	3.593	3.599	3.634	3.566
3	3.255	2.618	3.599	3.594
4	4.050	3.780	3.273	3.752
5	2.801	3.274	4.001	3.778
6	2.436	1.921	1.921	1.921
7	1.892	2.193	2.184	3.588
8	1.950	2.164	2.153	2.674
9	3.906	3.172	3.274	3.273
10	3.292	3.750	3.558	3.571

11	4.275	4.630	3.880	4.133
12	3.407	4.245	1.922	2.899
13	4.214	3.841	4.034	4.010
14	3.052	3.599	3.274	3.273
15	3.293	3.653	3.583	2.880
16	3.966	4.044	4.043	3.512
17	4.823	5.456	5.456	5.650
18	3.821	4.116	2.896	3.823
19	4.327	4.322	5.378	4.614
20	4.376	4.075	5.005	4.161
21	4.431	4.634	4.529	4.555
22	4.886	4.404	5.792	4.380
23	3.782	3.893	4.118	3.522
24	5.443	5.146	5.336	4.961
25	5.823	5.649	5.117	5.263
26	5.853	5.646	5.265	5.094
27	5.468	5.633	5.124	5.318
28	5.726	5.064	5.635	5.160

Table 6- Comparison of predicted activities of various kNN-MFA method for test set of compounds

Com.	**Actual Activity***	**Predicted Activity by kNN-MFA with SW**	**Predicted Activity by kNN-MFA with SA**	**Predicted Activity by kNN-MFA with GA**
9	2.705	2.887	3.015	3.332
19	5.823	5.661	5.143	5.135
20	4.318	5.133	5.129	3.350

*** = Observed activity in mmol**

It is known that the CoMFA method provides significant value in terms of a new molecule design, when contours of the PLS coefficients are visualized for the set of molecules. Similarly, the kNN-MFA models provide direction for the design of new molecules in a rather convenient way. The points which contribute to the SA kNN-MFA models in all three data sets are displayed in Figures 1-3. The range of property values for the chosen points may aid in the design of new potent molecules (Figures 1-3). The range is based on the variation of the field values at the chosen points using the most active molecule and its nearest neighbor set.

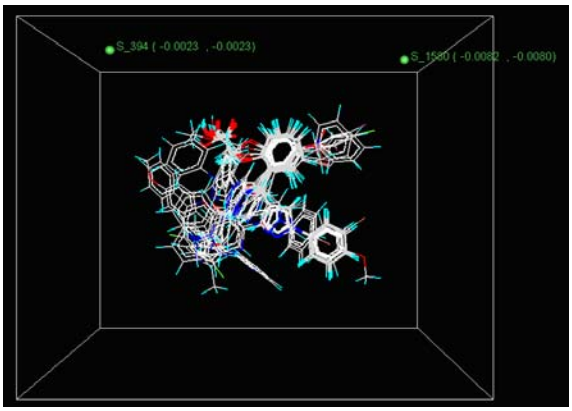


Fig. 1 - Distribution of chosen points in the SW kNN-MFA method

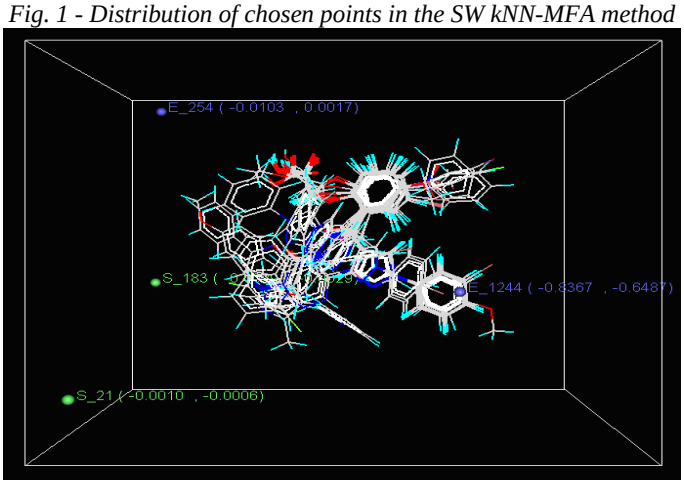


Fig. 2 - Distribution of chosen points in the SA kNN-MFA method

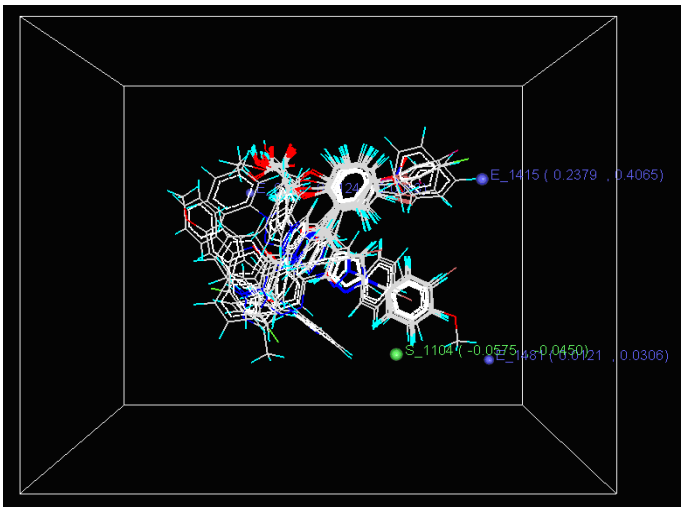


Fig. 3 - Distribution of chosen points in the GA kNN-MFA method.

The q^2, pred_r^2, Vn and k value of kNN-MFA with SW, SA & GA were (0.8392, 0.7059, 2/2) (0.6725, 0.6716, 2/4) and (0.6832, 0.6716, 2/4) although there are no common descriptors among these three methods, SW kNN-MFA method have better q2 (0.8392) and pred_r2 (0.7059) than other two methods, model validation correctly predicts activity 83.9% and 70.5% for the training and test set respectively. It uses 2 steric descriptors with 2 k nearest neighbor to evaluate activity of new molecule, So model generated by SA kNN-MFA are best model.

S_1580, S_394 in SW kNN-MFA, S_21,S_183 in SA kNN-MFA and S_1104 are steric field descriptors similarly E_1244, E_254 in SA kNN-MFA and E_613, E_1481, E_1415 are electrostatic field descriptors.

Negative value in electrostatic field descriptors indicates that negative electronic potential is required to increase activity and more electronegative substituents group is preferred in that position, positive range indicates that group that imparting positive electrostatic potential is favorable for activity so less electronegative group is preferred in that region.

Negative range in steric descriptors indicates that negative steric potential is favorable for activity and less bulky substituents group is preferred in that region, Positive value of steric descriptors reveals that positive steric potential is favorable for increase in activity and more bulky group is preferred in that region.

4. Conclusions

The proposed models, due to the high internal and external predictive ability, can therefore act as a useful aid to the costly and time consuming experiments for determining the molar concentration of a compound required to achieve better $CRTh_2$ antagonist activity. Our results lead to the conclusion that the $CRTh_2$ antagonist activities of 2,4-disubstituted-phenoxy acetic acid derivatives can be increased if substituents that bring about changes in the molecule as mentioned above are attached to it.

Acknowledgement

One of the authors Abhishek K. Jain would like to thank AICTE, New Delhi for providing scholarship.

References

[1] R. A. Lewis, N. A. Soter, P. T. Diamond, K. F. Austen, J. A. Oates, L. J. Roberts, J. Immunol, **129**, 1627 (1982)

[2] A. N. Hata & R. M. Breyer, Pharmacol Ther, **103**, 147 (2004)

[3] H. Sugimoto, M. Shichijo, T. Iino, Y. Manabe, A. Watanabe, M. Shimazaki, F. Gantner, & K. B. Bacon, J Pharmacol Exp Ther., **305***, 347* (2003)

[4] M. J. Robarge, D. C. Bom, L. N. Tumey, N. Varga,, E. Gleason, D. Silver, J. Song, S. M. Murphy, G. Ekema, C. Doucette, Bioorg Med Chem Lett., **15**, 1749 (2005)

[5] T. Ulven, & E. Kostenis, J Med Chem., **48**, 897 (2005)

[6] A. K. Jain, V. Ravichandran, R. Singh, S. Sharma, V. K. Mourya, R. K. Agrawal, Digest Journal of Nanomaterials and Biostructures, **3**, 63 (2008)

[7] V. Ravichandran, & R. K. Agrawal, Bioorganic Medicinal Chemistry Letter, **17**, 2197 (2007a)

[8] V. Ravichandran,, P. K. Jain,, V. K. Mourya, & R. K. Agrawal, Medicinal Chemistry Research , (in press), (2007b)

[9] V. Ravichandran, V. K. Mourya, & R. K. Agrawal, Arkivoc, **XIV**, 204 (2007c)

[10] V. Ravichandran, V. K. Mourya, & R. K. Agrawal, Internet Electronic Journal of Molecular Design, (Inpress), (2007d)

[11] K. K. Sahu, V. Ravichandran, P. K. Jain, S. Sharma, V. K. Mourya & R. K. Agrawal, Acta Chimica Slovenica, (Inpress),(2007b)

[12] K. K. Sahu, V. Ravichandran, V. K. Mourya & R. K. Agrawal, Medicinal Chemistry Research, **15**, 418 (2007a)

[13] R. D. Cramer, D. E. Patterson, J. D. Bunce, J. Am. Chem. Soc., **110**, 5959 (1988)

[14] G. Klebe, U. Abraham, T. Mietzner, J. Med. Chem., **37**,24 (1994)

[15] M. Baroni, G. Costantino, G. Cruciani, D. Riganelli, R. Valigi, S. Clementi, An Advanced Chemometric Tool for Handling 3D-QSAR Problems, Quant. Struct.-Act. Relat., **12**, 9 (1993)

[16] S. J. Cho & A. Tropsha, J. Med. Chem.,**38**, 1060 (1995)

[17] A. Golbraikh & A. Tropsha, J. Chem. Inf. Comput. Sci., **43**, 144 (2003)

[18] M. A. Sharaf, D. L. Illman, B. R. Kowalski, Chemometrics, Wiley, New York, (1986)

[19] J. Holland, Adaptation in Natural and Artificial Systems, University of Michigan Press, (1975)

[20] W. Zheng & A. Tropsha, J. Chem. Inf. Comput. Sci., **40**, 185 (2000)

[21] N. Gilbert, Statistics; W. B. Saunders, Co., Philadelphia, PA (1976).

[22] VLifeMDS2.0; Molecular Design Suite, Vlife Sciences Technologies Pvt. Ltd., Pune, India, (2004), (www.vlifesciences.com)