Available online at www.scholarsresearchlibrary.com

Scholars Research Library

Der Pharmacia Lettre, 2010: 2 (1) 150-161
(http://scholarsresearchlibrary.com/archive.html)

Scholars Research Library
ISSN 0975-5071
USA CODEN: DPLEB4

QSAR and k-Nearest Neighbour Molecular Field Analysis (k-NN MFA) Classification Analysis of Studies of Some Bemzimidazoles Derivatives Antibacterial activity Against *Escherichia coli*

M. C. Sharma*[a], Smita Sharma[b], D. V. Kohli[c], S. C. Chaturvedi[d]

[a] *School of Pharmacy, Devi Ahilya Vishwavidyalaya, Indore (M.P), India*
[b] *Department of Chemistry, Yadhunath Mahavidyalya, Bhind (M.P), India*
[c] *Department of Pharmaceutical Sciences, Dr. Hari Singh Gour University, Sagar (M.P), India*
[d] *Shri Arvindo Institute of Pharmacy, Ujjain Road, Indore (M.P), India*

Abstract

A Quantitative Structure Activity Relationship study on a Series of 17 molecules of (benzimidazole compounds) with antimicrobial activity analogues was made using combination of various physicochemical descriptors. (Thermodynamic, electronic and spatial). Several statistical expressions for 2D QSAR & 3D QSAR were developed using stepwise partial least square (PLS) regression analysis and K-Nearest neighboring molecular field analysis (K-NN-MFA) respectively. The best Quantitative Structure Activity Relationship models were further cross validated. The study revealed that the alignment independent descriptors contributed positive and path count contributed negatively in 2D-QSAR analysis and Electrostatic descriptor contributed positive and steric descriptor contributed negatively in 3D QSAR analysis. 2D-QSAR model developed using partial least square regression approach. Negative logarithmic value of (–PMIC) was taken as dependent variable and T_N_N_4, T_2_C_1, T_T_C_4, T_T_S_7 T_2_C_1, ChiV3, T_O_O_7 was taken as independent varable. The analysis resulted in the following 2D-equation suggest that, BA = [-1.2083 (± 9.17986)] + T_O_O_7 1 [0.152608(± 0.336984)] + T_2_C_1 [0.1510 (± 0.000191822)] + T_T_C_4 [-1.8960 (± 0.504224)] n=11, r^2 = 0.9717, q^2 = 0.8367, F test = 60.0062, pred_r^2 = 0.6547, a lipophillic group, which is less bulkier at Ar, is important for guiding the design of a new molecule.3D-QSAR model developed using K-nearest neighbour method (training set =11 and test set = 6). Out of several model were developed. The best model derived by the method have cross-validated coefficient q^2 value is 0.7926, Predict r^2 value is 0.8919, k Nearest Neighbor is 2, Degree of freedom = 6. The steric and electrostatic descriptors at the grid points, E_442, S_473, S_135, S_190 plays important role for design of new molecule. QSAR analysis of series of benzimidazole compounds informed that electronegative and less bulky group increases the biological activity.

Key Words: - 2D-QSAR (PLS), 3D-QSAR (k-NN-MFA), *E. coli.*

Introduction

The emergence and spread of antimicrobial resistance has become one of the most serious public health concerns across the world. Antimicrobial resistance refers to micro-organism that have developed the ability to inactivate, exclude or block the inhibitory or lethal mechanism of the antimicrobial agents [1-4]. Benzimidazole Compounds constitute an important class of heterocyclic aromatic organic compounds for their versatile pharmacological activities such as antibacterial, antifungal, antihelmintic, antiallergic, antineoplastic, local analgesic, antihistaminic, vasodilative, hypotensive, and spasmolytic activities [5-6]. In the present study, 2D-QSAR, 3D-QSAR analysis of some bemzimidazole compounds with antimicrobial activity was performed by using Partial least square regression (PLS) and k-nearest neighbour method (KNN) approach. A data set of 17 molecules was taken from K.F.Ansari *et al* [7] and MIC value of molecules were converted to negative logarithmic values (–PMIC) by using software VLIFE MDS 3.5 [8].

Materials and Methods

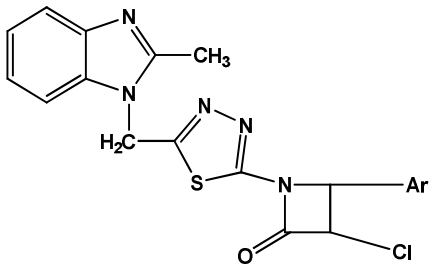
The dataset consist of structurally diverse compounds reported for Antimicrobial inhibitory activities. The selected series comprises of seventeen Benzimidazole analogues reported by K.F.Ansari et al [7] (Table 1).The Antimicrobial activity of compounds in the series is reported as pIC_{50} values where IC_{50} refers to experimentally determined concentration required to inhibit 50% of Antimicrobial activity. The compounds in the selected series were randomly divided into two sets with 11 compounds used as a training set in developing regression models and the remaining 6 as validation set (Test set) in the prediction of biological activity.

2D-QSAR Methodology

The molecular structures of the compounds in selected series were sketched using V-life MDS 3.5 of V-Life sciences molecular modeling software. The sketched structures were then transferred to three dimensional structures (3D). The geometries of generated 3D structures were optimized using MMFF94 force field as implemented in the V-Life MDS3.5. The gradient norm 0.001 kcal/Å was used to calculate electronic, geometric and energetic parameters for the isolated molecules. The optimized geometries of the molecules were used to compute the necessary quantum chemical descriptors available in the V-Life MDS 3.5. Further, calculated some selected molecular descriptors available in the software V-Life MDS3.0 Variable-selection for the QSAR modeling was carried out by stepwise partial least square regression method (PLS) using statistical program of V-Life MDS 3.5. The program employs a stepwise technique, *i.e.*, only one parameter at a time was added to a model and always in the order of most significant to least significant in terms of F-test values. Statistical parameters were calculated subsequently for each step in the process, so the significance of the added parameter could be verified. The goodness of the correlation is tested by the regression coefficient (r^2), the F-test and the standard error of estimate (SEE). The correlation coefficient values closer to 1.0 represent the better fit of the model. The F-test reflects the ratio of the variance explained by the model and the variance

due to the error in the model (i.e., the variance not explained by the model). High values of the F-test indicate that the model is statistically significant. Finally, the derived QSAR models were used for the prediction of the activity values of the compounds in the test set and the external validation parameter, predictive r2 (r^2pred) was calculated for evaluating the predictive capacity of the model. A value of r^2_ pred. greater than 0.5 indicates the good predictive capacity of the QSAR model. The different statistical models were developed by using this method (table-3).

Table-1



S.NO.	COMPOUND CODE NO.	STRUCTURE	MIC
1	5a	-C_6H_5	100
2	5b	4-Br- C_6H_4	500
3	5c	4-Cl- C_6H_4	200
4	5f	4-OCH3- C_6H_4	100
5	5g	2-CH3- C_6H_4	6.25
6	5i	2-OH C_6H_4	12.5
7	5k	4-OH C_6H_4	25
8	5l	4-NH2 C_6H_4	50
9	6a	C_6H_5	64
10	6b	2-Cl C_6H_4	6.3
11	6e	4-OCH_3 C_6H_4	100
12	6f	2-OCH_3 C_6H_4	100
13	6h	4-CH_3 C_6H_4	3.2
14	6i	2-OH C_6H_4	0.8
15	6j	3-OH C_6H_4	1.6
16	6k	4-OH C_6H_4	50
17	6l	4-NH2 C_6H_4	200

The models showed the better correlation between biological activity and physicochemical descriptor values. Two Dimensional Structures of 17 benzimidazole compounds were used for further studies before starting any computational methods energy was minimized. Total of 277 descriptors were calculated and these descriptors were further classified into following groups' viz. constitutional, topological, geometrical, electrostatic, quantum-chemical and thermo-dynamical descriptors. Partial least squares regression (PLSR) methodologies were used to yield the QSAR equation and to validate the predictive ability of the model using test set. The

selection of descriptors is based on F-test value and r^2 and minimum of one-parameter correlation. Partial least square regression (PLS) analysis was used to derive a linear correlation between the descriptors value (independent variables) and the inhibitory activity values (dependent variables) the cross validation analysis was performed using leave-one-out (LOO) method in which one compound was removed from the dataset and its activity was predicted using the model built from rest of the dataset. The different statistical models were developed by using this method. The models showed the better correlation between biological activity and physicochemical descriptor values in (Table- 3). The values of descriptor were used to create a 2D-model for assuming the biological activity with the help statistical methods. The values of descriptor are given in (Table- 4). All the data and graphs indicate biological activities of molecules are correlated with physicochemical properties in (Fig 2-3).

Table-2. Training and Test set of Series of Benzimidazole Compounds

Sr No.	Molecule	Structure AR	MIC	-PMIC Value
1	5a	-C_6H_5	100	4
2	5b	4-Br-C_6H_4	500	3.30103
3	5c*	2-Cl-C_6H_4	200	3.69897
4	5f	4-OCH_3-C_6H_4	100	4
5	5g	2-CH_3-C_6H_4	6.25	5.20412
6	5i*	2-OH C_6H_4	12.5	4.90309
7	5k	4-OH C_6H_4	25	4.60206
8	5l*	4-$NH_6C_6H_4$	50	4.30103
9	6a	-C_6H_5	64	4.19382
10	6b	2-Cl C_6H_4	6.3	5.200659
11	6e*	4-$OCH_3C_6H_4$	100	4
12	6f	2-$CH_3C_6H_4$	100	4
13	6h*	4-$CH_3C_6H_4$	3.2	5.49485
14	6i	2-OH C_6H_4	0.8	6.09691
15	6j	3-OH C_6H_4	1.6	5.79588
16	6k	4-OH C_6H_4	50	4.30103
17	6l	4-NH_2 C_6H_4	200	3.69897

* Test compound-

3D-QSAR Methodology

Alignment

The template based alignment method was used to aligning all the structure of benzimidazole compounds depicted in (Fig-4).Many topological descriptors can be used to describe organic molecular structure with QSAR aims. Resent trends in 2D, 3D QSAR have focused on the development of procedure that allows selection of optimal variables from the pool of descriptors of chemical structures i.e. ones that are most meaningful and statistically significant in terms of correlation with biological activity. This is accomplished by combining one of the stochastic search methods such as SA, GAs, or evolutionary algorithms with the correlation methods such as MLR, PLSR, or artificial neural networks [9-14]. The k-NN MFA, used for 3D QSAR

analysis of the present data set adopts a k-nearest neighbour principle for generating relationships of molecular fields with the experimentally reported activity. The variables and optimal k values were chosen using three variable selection methods viz. SW, SA, and GA. Like many 3D QSAR methods, k-NN 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 are generated using the sphere exclusion algorithm. This algorithm allows the construction of training sets covering descriptor space occupied by representative points. Once the training and test sets are generated, k-NN methodology is applied to the descriptors generated over the grid [15].

Calculation of fields

Insert the molecules and its biological activity values (-PMIC value) under the 3D-QSAR worksheet both molecules and -PMIC values were take in separate column and check the proper alignment of molecules before the 3D QSAR studies. Alignment play crucial role for computing the fields. The steric and electrostatic descriptors at the grid points, E_442, S_473, S_135, S_190 plays important role for design of new molecule depicted in (Fig-5, 6, 7). Blue spheres under the grid indicate the electrostatic involvement in the aligned molecules. Positive range indicates that positive electrostatic potential is favorable for increase in the activity and hence a electronegative substituent group is preferred in that region. Green spheres under the grid indicate the steric involvement in the aligned molecules. Positive range indicates that positive steric potential is favorable for increase in the activity and hence less bulky substituent group is preferred in that region.

Simulated Annealing K-NN QSAR Algorithm

k-Nearest Neighbour method of Molecular field analysis (KNN-MFA) is the 3D-QSAR method it is like COMFA (comparative molecular field analysis) which has been used to produce the three dimensional models to indicate the regions that affect biological activity with a change in the chemical substitution. This method uses simulated annealing variable selection and k-NN principle to build QSAR model. The best model derived by the method have cross-validated coefficient q^2 value is 0.7926, Predict r^2 value is 0.8919, k Nearest Neighbor is 2, Degree of freedom = 6.

k-NN MFA 3D-QSAR MODELS

To derive the kNN-MFA descriptor fields, a 3D cubic lattice grid in *x*, *y* and *z* directions, was created to encompass the aligned molecules. kNN-MFA descriptors were calculated using an sp3 carbon probe atom with a van der Waals radius of 1.52 Å and a charge of +1.0 to generate steric field energies and electrostatic fields with the distance dependant dielectric at each lattice point. The steric and electrostatic energy values were truncated at a default value of 30 kcal/mol.

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 Batsman 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.

k-NN-MFA with Genetic Algorithm

Genetic algorithms (GA) first described by Holland 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. The 3D QSAR for molecular field analysis was performed using the k Nearest Neighbour MFA method using software V-LIFE MDS 3.5.

Randomization Test.

To evaluate the statistical significance of the QSAR model for an actual data set, we have employed a one-tail hypothesis testing. 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 Z_{score} [16-17].

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 neighbours; *q*2, cross validated *r*2 (by the leave-one-out method); pred_r2, predicted *r*2 for the external test set; Z score, the Z score calculated by *q*2 in the randomization test; best_ran_q2, the highest *q*2 value in the randomization test; and R, the statistical significance parameter obtained by the randomization test.

Results and Discussion

2D-QSAR

The total set of Benzimidazole molecules (17 compounds) was divided into training (11 compounds) and test (6 compounds) sets in the approximate ratio 70:30. In the partial least square regression method, several models were generated for the given or selected members of training and test sets. The different statistical models were developed by using this method. The models showed the better correlation between biological activity and physicochemical descriptor values. The correlation coefficient (r^2 value) and cross validated squared correlation coefficient value (q^2) was found to be r2 = 0.9717, q2 = 0.8367 (model-I), r2 = 0.8553 q2 = 0.7986, (model-II) r2 = 0.9121 q2 = 0.5755 (model-III), r2 = 0.9166 q2= 0.7538 (model-IV), r2=0.9165 q2 = 0.7367 (model-V), r2 = 0.8171 q2= 0.5672 (model-VI).

Table-3-Statistical analysis report of 2D QSAR by PLS methods

S.No	Parameters	Model-I	Model-II	Model-III	Model-IV	Model-V	Model-VI
1	Optimum components	2	2	2	2	2	2
2	n	11	11	11	11	11	11
3	r^2	0.971	0.8553	0.9121	0.9166	0.9165	0.8171
4	q^2	0.836	0.7386	0.5755	0.7538	0.7367	0.5672
5	F-test	60.02	59.0849	44.1201	35.166	41.7200	49.1456
6	r^2se	0.142	0.2692	0.4223	0.4197	0.3889	0.5349
7	q^2se	1.002	0.7820	0.9282	0.7211	0.5439	0.3661
8	Pred_r^2	0.822	0.6932	0.7812	0.8154	0.7284	0.7929
9	Pred_q^2	0.450	0.2590	0.2674	0.5366	0.4896	0.2958
10	Degree freedom	7	10	8	**8**	8	8

n = number of molecules, r2 = correlation coefficient, q2 =cross validated r^2 values-Test = test of significance, r2se = standard error of r^2, q2se = stander error of q^2, Pred_ r^2 = predicted r^2.

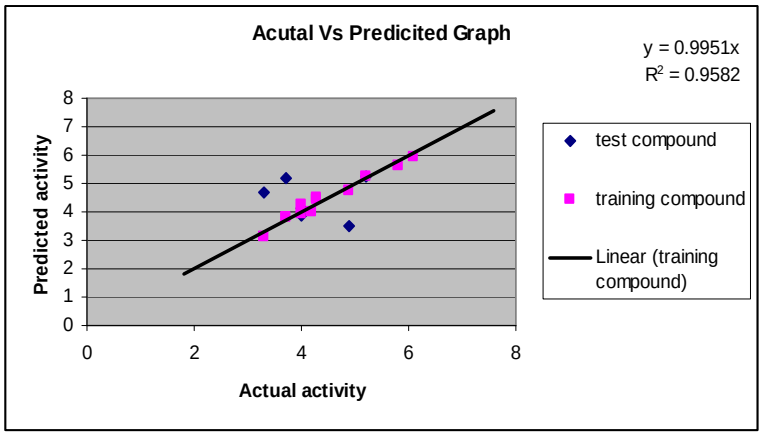
Table:-4 Calculated value of descriptors for the given series of compounds

S.No	Mol.Wt.	T_O_O_ 7	T_N_N_4	T_2_C_1	T_T_C_4	T_T_S_7	Chi V3
5a	409.89	5	2	34	37	14	2.925015
5b	488.79	4	1	35	32	22	3.061698
5c	444.34	4	3	37	32	28	3.345095
5f	439.54	3	3	39	35	29	3.568182
5g	423.92	8	1	34	43	18	3.104008
5i	425.89	5	2	34	41	16	3.09536
5k	484.81	8	1	34	37	14	2.961535
5l	424.32	4	1	34	37	14	2.907311
6a	409.89	3	0	34	37	14	2.884542
6b	444.34	5	2	37	43	20	4.917457
6e	439.92	6	0	39	40	28	5.1707
6f	423.57	6	1	41	40	32	5.454097
6h	325.23	5	1	43	43	33	5.677184
6i	583.89	5	4	34	45	30	5.927184
6j	214.11	5	3	35	47	30	6.177184
6k	321. 82	5	2	37	49	33	6.427184
6l	132.22	3	1	39	35	12	2.808455

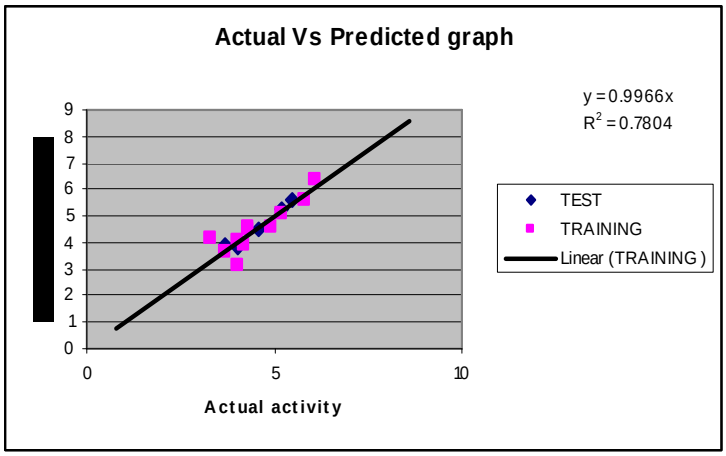
Table -5 Predicted activities of different models by PLS methods

Molecules	Act. Activity	Predicted activity of different models by PLS methods					
	PIC_50	Model-I	Model-II	Model-III	Model-IV	Model-V	Model-VI
5a	4	3.96858	4.145579	4.27702	3.911395	4.183391	4.102102
5b	3.30103	3.126616	3.175735	3.620945	3.225584	2.992704	3.185116
5c	3.69897	3.802313	3.610836	3.820945	3.425584	3.550554	3.641605
5f	**4**	**4.137606**	**3.830836**	**3.949888**	**4.107766**	**4.170344**	**3.941605**
5g	5.20412	5.23571	5.113474	5.28421	5.452134	5.106625	5.450295
5i	4.90309	4.73407	4.596427	4.769471	4.852134	4.985048	4.549847
5k	**4.60206**	**4.69671**	**4.479408**	**4.78422**	**4.571395**	**4.655672**	**4.51384**
5l	4.30103	4.429671	4.479408	3.98336	4.13031	4.37802	4.219681
6a	4.19382	4.01858	3.945579	3.99817	4.02031	4.24801	3.93643
6b	**5.200659**	**5.168747**	**5.264916**	**5.179064**	**5.40031**	**5.11809**	**5.14606**
6e	**4**	**3.906043**	**3.96359**	**3.99655**	**4.149956**	**4.264314**	**4.203758**
6f	4	4.26681	4.11201	4.213655	4.149956	3.922164	4.197247
6h	**5.49485**	**5.25963**	**5.591001**	**5.642598**	**5.232138**	**5.341954**	**5.597247**
6i	6.09691	5.91395	6.391001	6.47154	6.51432	6.368584	6.740758
6j	5.79588	5.648235	5.591001	5.67154	5.91432	5.853046	5.640758
6k	4.30103	4.517238	4.591001	4.142598	4.232138	4.196264	4.440758
6l	**3.69897**	**3.482712**	**3.904751**	**3.786356**	**3.875938**	**3.538199**	**3.790364**

Model .1 Plot between Actual activity and predicted activity



Model.2 Plot between Actual activity and predicted activity



K-Nearest Neighbor method of Molecular field analysis

A set of 17 molecules of benzimidazoles (11 molecules as a training set and 6 molecules as a test set for 3D-QSAR analysis). K-Nearest Neighbor method of Molecular field analysis (KNN-MFA) is the 3D-QSAR method it is like COMFA (comparative molecular field analysis) which has been used to produce the three dimensional models to indicate the regions that affect biological activity with a change in the chemical substitution. This method uses simulated annealing variable selection and k-NN principle to build QSAR model. Models with good statistical qualities were developed using the software VLIFE MDS 3.5. Out of several models developed, the best 3D-QSAR model having highest cross validated squared correlation coefficient was q^2 value is 0.7926, Predict r^2 value is 0.8919.

Table-6 3D QSAR statistical data of different models by PLS method

S.N.	PARAMETERS	MODEL-I	MODEL-II	MODEL-III	MODEL-IV	MODEL-V	MODEL-VI
1	**Optimum components**	4	4	4	4	4	4
2	**n**	17	17	17	17	17	17
3	**r^2**	0.9111	0.8861	0.7601	0.6942	0.8181	0.8649
4	**q^2**	0.7694	0.6352	0.6937	0.5809	0.7153	0.7934
5	**F-test**	73.4033	56.1814	43.0757	63.2857	97.0090	49.4720
6	**r^2se**	0.2347	0.2067	0.1597	0.1371	0.1581	0.2173
7	**q^2se**	0.4814	0.3619	0.4097	0.5702	0.5201	0.4661
8	**Pred_r^2**	0.7356	0.6457	0.7422	0.6670	0.8061	0.7292
9	**Pred_q^2**	0.4508	0.2590	0.2674	0.5366	0.4896	0.2958
10	**Degree of freedom**	9	9	8	8	8	8

Table-7 Statistical data of K-NN-MFA method of 3D-QSAR

Parameters	Steric field	Electrostatic field
k Nearest Neighbour	4	4
Degree of freedom	22	24
n	17	17
q2	0.8926	0.7980
q2_se	0.4566	0.3201
pred_r^2	0.7919	0.6914
pred_r^2se	0.5656	0.6413

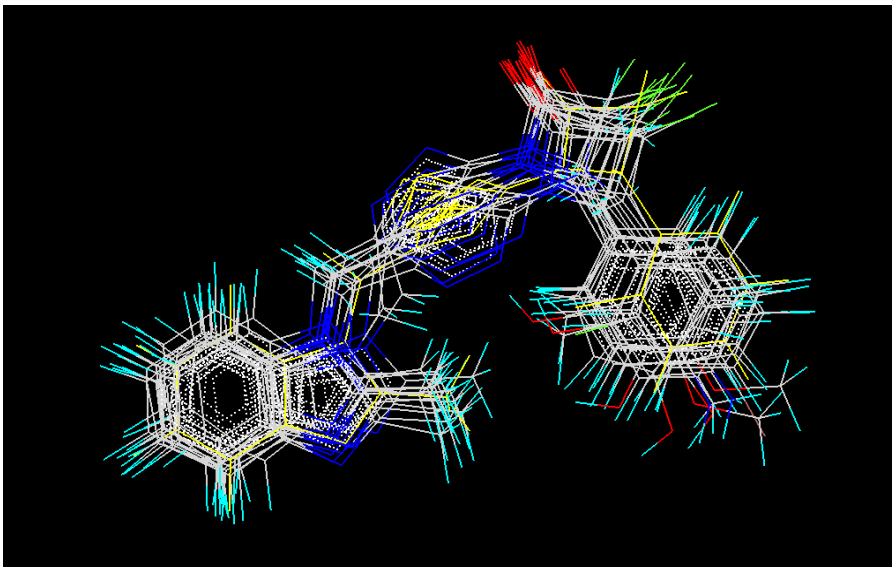


Figure 4: Alignment of benzimidazole compounds on the basis of template

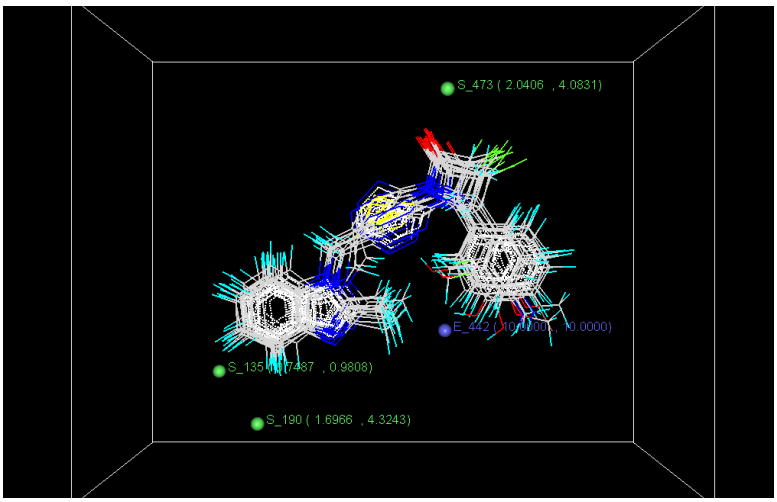
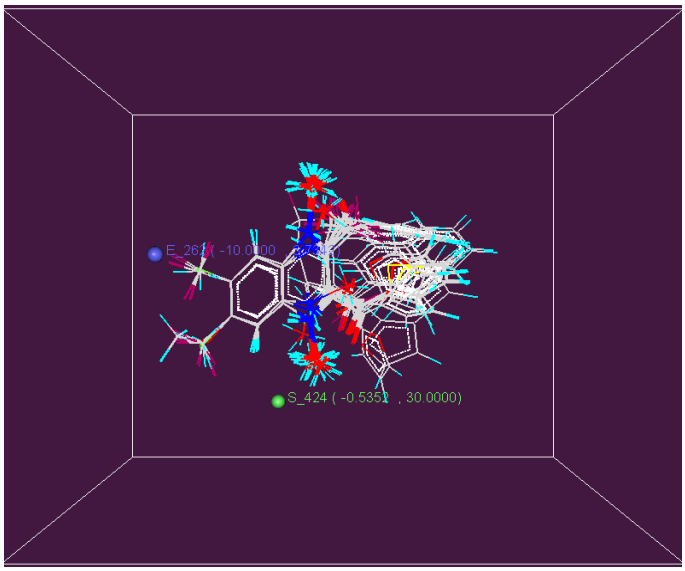


Figure-5 Model-1



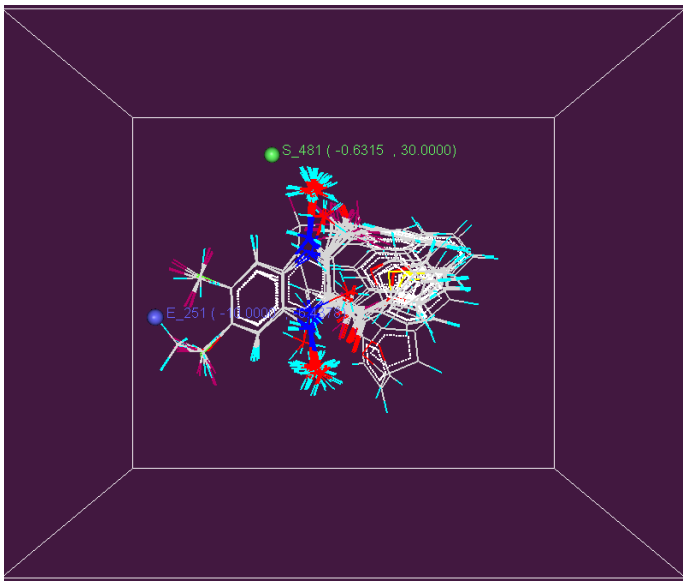


Figure: 6 **Figure: 7**

Blue spheres (electrostatic field) green spheres (static field)

Conclusion

A set of 17 compounds of Benzimidazole were subjected to 2D and 3D QSAR analysis using partial least square regression and k nearest neighbors molecular field analysis method respectively to design benzimidazole compounds as antimicrobial agents. This study informed that electronegative and less bulky group will increase the antimicrobial activity of benzimidazole compounds.

Acknowledgement

The authors are thankful to Mr.Amit Badi for Cooperating and given valuable suggestions and Vlife Science Technologies Pvt. Ltd, 1, Akshay 50, Anand Park, Aundh, Pune, India to provide trial version of software.

References

[1] Tolaro., Foundation of Microbiology, W.C. Brown Publisher, Dubuque; **1993**, pp.326.
[2] Tortora.,Microbiology an Introduction, 7th edition, Addison Wisley Longma Publication, San Franciso., 2001, pp.19.
[3] Purohit S.S., Microbiology fundamentals and applications, 6th, Agrobioas Ltd, India; 2003, pp.505.
[4] Kasper D.L.,Harrison Principles of Internal Medicine, 10th edn, MeGraw-Hill Medical Publishing Division, New York; **2005,** pp.953.
[5] Saleem K., Khan S.A., Singh N., *Eur. J. Med Chem*., **2007,** 1 -6.
[6] *Block H., Wilson and Gisvold's Textbook of Organic Medicinal and Pharmaceutical Chemistry, Eleventh edition* Lippincott-Raven Publisher., ***2004***,*265* –275.

[7] Ansari K.F., *Bioorg.Med.Chem*., **2008,**1-6.
[8] V-life MDS3.5 Vlife science technologies pvt. Ltd, 1 akshay residency, plot 50 anand park, aundh pune 411007
[9]SutterJ.M.,DixonS.L.,JursP.C.,*J.Chem.Inf.Comput.Sci*.,**1995**,35,77-84.
[10] Rogers D., Hopfinger A.J., *J.Chem. Inf. Comput. Sci*., **1994**, 34,854-866.
[11]Kubinyi.H.,*Quant.Struct.Act.Relat*.,**1994**,13,285-294.
[12] Kubinyi.H., *Quant. Struct.Act. Relat*., **1994,**13,393-401.
[13] Luke B.T., *J. Chem. Inf. Comput. Sci*., **1994,** 34,1279-1287.
[14] So. S.S., Karplus M., *J. Med. Chem*., **1996,** 39, 1521-1530.
[15] Ajmani S., Jadhav K., Kulkarni S.A., *J.Chem. Inf. Model*., **2006**, 46, 24-31.
[16] Sharaf M.A., Illman D.L., Kowalski B.R., Chemometrics, Wiley, New York, **1986.**
[17] Holland J., Adaptation in Natural and Artificial Systems, University of Michigan Press, **1975**.