EXCELLENT PUBLISHERS

International Journal of Current Research and Academic Review

ISSN: 2347-3215 Volume 3 Number 5 (May-2015) pp. 288-307

www.ijcrar.com

Analysis of Naphthoquinone Derivatives as Topoisomerase I Inhibitors using Fragment Based QSAR

Bastikar Virupaksha[1,2*] and Khadke Prashant[1]

[1]Department of Bioinformatics, JJT University, Rajasthan, India
[2]Department of Structural Biology, University of Rome Tor Vergata, Rome, Italy
Corresponding author

KEYWORDS

Drug Design;
Fragment based QSAR;
group based QSAR;
k-nearest neighbor;
Naphthoquinones;
Partial Least Square;
SAR;
Topoisomerase I;
topological descriptors

ABSTRACT

In this study an attempt was made to understand the structural requirements for Topoisomerase I (Topo I) inhibition using a novel Group based QSAR (GQSAR) or fragment based QSAR technique. Here we combined the GQSAR technology with conventional 2D and 3D QSAR to derive GQSAR models for various reported Naphthoquinone derivatives. Various regression models such as Multiple Regression (MRA), Partial Least Square (PLS) and Principal Component Analysis (PCA) as well as k-Nearest neighbor (k-NN) QSAR were used to develop several combined 2D and 3D GQSAR models. The GQSAR analyses revealed the importance of Geometrical topological indices and Baumann's alignment independent topological descriptors along with dipole moment and other general descriptors like HBonddonor and XYHydrophilic etc for governing the activity variation. Further the GQSAR showed that chemical variation like presence of substituted double bonded C atom separated from oxygen by 6 bonds and HBonddonor count are highly influential for achieving highly potent Topo I inhibitors. The Naphthoquinone derivatives having 2-CH(OX)-($CH_2CH=CMe_2$)-5,8-dihydroxy-1,4-naphthoquinone substitutions are most important fragments for the inhibitory activity. In addition the k-nearest neighbor classification model resulted in 3 important descriptors like moment of inertia, quadrapole and hydrogen count. The developed models are interpretable with good statistical and predictive significance and can be used for guiding ligand modification for development of potential new Topo I inhibitors. From the present study it can be seen that the substitutions made on 2-CH(OX)-($CH_2CH=CMe_2$)-5,8-dihydroxy-1,4-naphthoquinone position can result in better Topo I inhibitors.

Introduction

Naphthoquinones are wide-spread phenolic compounds in nature, based on the C6-C4 skeleton. 1, 4-Naphthoquinones can be viewed as derivatives of naphthalene through the replacement of two hydrogen atoms by two ketone groups (Figure 1a).

They are products of bacterial and fungal as well as high-plants secondary metabolism. Naphthoquinones display very significant pharmacological properties--they are cytotoxic, they have significant antibacterial, antifungal, antiviral, insecticidal, anti-inflammatory, and antipyretic properties. Pharmacological effects to cardiovascular and reproductive systems have been demonstrated too. The mechanism of their effect is highly large and complex--they bind to DNA and inhibit the processes of replication, interact with numerous proteins (enzymes) and disturb cell and mitochondrial membranes, interfere with electrons of the respiratory chain on mitochondrial membranes [1].

Many derivatives of Naphthoquinones have been reported to show anti Topo I enzyme activity [2, 3-7]. It has been suggested that they inhibit the enzyme by binding to the Zn finger domain of the protein (Figure 1b) [8]. Attempts have been made to establish a Quantitative structure activity relationship of the naphthoquinone derivatives so as to obtain new better molecules having anti Topo I activity. A number of in-silico and experimental approaches have been mentioned for assisting in the design of novel and more effective naphthoquinone molecules as Topo I inhibitors. Many 2-D QSAR models have been developed to relate the structure of naphthoquinone derivatives with their biological activity. However, they mainly focus on a particular chemical class of molecules. This paper introduces a novel approach known as Group QSAR (GQSAR) or fragment based QSAR to gain deeper insights into the structural requirements for Topoisomerase I inhibition and develop quantitative models for the development of new naphthoquinones. GQSAR is a recent QSAR method developed, which addresses the challenges of QSAR model interpretation and the inverse QSAR problem [9]. GQSAR method comprises of three steps: (1) generation of molecule fragments using a set of predefined chemical rules, (2) calculation of descriptors for the generated fragments, (3) build statistical models using the calculated fragment descriptors and their interactions. GQSAR thus allows establishing a correlation of chemical group/fragment variation at different molecular sites of interest with the biological activity. Fragmentation is done by applying specific chemical rules for breaking the molecules along specific bonds and/or bonds on ring fusion and/or any pharmacophoric feature such as hydrogen bond acceptor, hydrogen bond donor, hydrophobic group, charged group etc. Thus, the GQSAR method deals with molecular fragments instead of the molecule as a whole. The fragment descriptors and their interactions are related to biological activity, resulting in model(s) that highlight important substitution site(s) along with their chemical nature and interactions. The suggested important fragments can be used as the building blocks to design novel molecules [16].

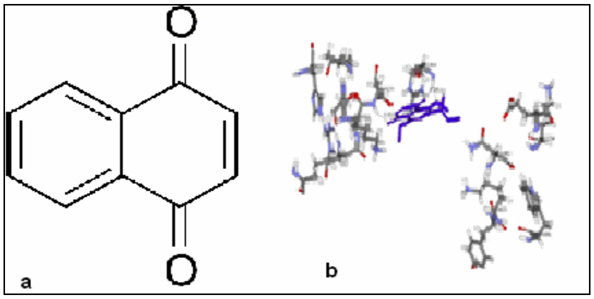
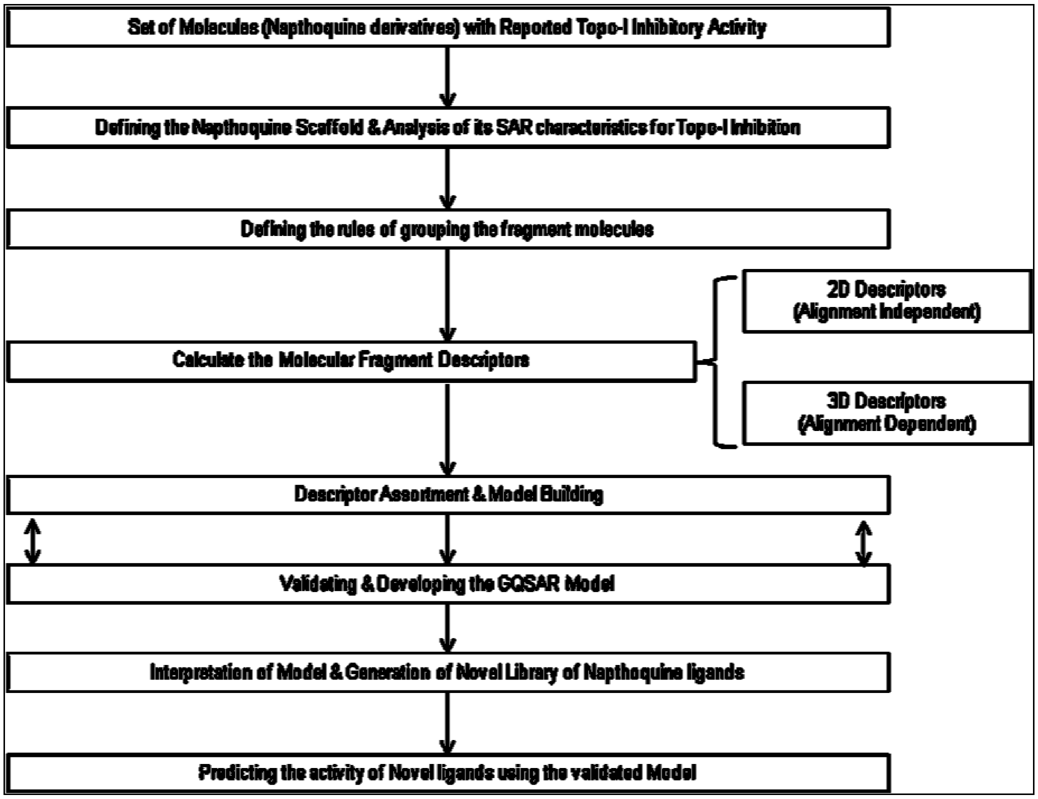


Figure 1- (a) Naphthoquinone Scaffold, (b) Naphthoquinone- Topo I complex

Methodology

All computations and molecular modeling studies were carried out on a Windows workstation using the molecular modeling software package VLife Molecular Design Suite (VLifeMDS) version 3.5. The schematic representation of the entire methodology is demonstrated in Figure 2.

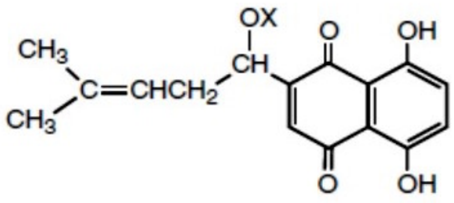
Figure.2 Flowchart of GQSAR


Set of Molecules (Napthoquine derivatives) with Reported Topo-I Inhibitory Activity
Defining the Napthoquine Scaffold & Analysis of its SAR characteristics for Topo-I Inhibition
Defining the rules of grouping the fragment molecules
Calculate the Molecular Fragment Descriptors
2D Descriptors (Alignment Independent)
3D Descriptors (Alignment Dependent)
Descriptor Assortment & Model Building
Validating & Developing the GQSAR Model
Interpretation of Model & Generation of Novel Library of Napthoquine ligands
Predicting the activity of Novel ligands using the validated Model

Dataset- A total of 90 naphthoquinone based inhibitors with corresponding biological activities, reviewed from various literature sources were used in the study (Table 1-6) [3, 5, 7, 10, 11]. The minimum inhibitory concentration (IC_{50}) values were converted to the corresponding pIC_{50} ($-\log IC_{50}$) values and used as dependent variables for the combined 2D and 3D GQSAR analysis. The pIC_{50} values span a range of 3-4 log units, providing a broad and homogenous data set for the GQSAR study. The initial structures of 90 compounds were constructed using the Marvin Sketch 5.3.8. These structures were incorporated into the VLife MDS GQSAR module.

Energy Minimization- Energy minimization of the ligand structures was performed using the Merck Molecular Force Field (MMFF), with a distance-independent dielectric constant of 1.0 and MMFF charges, with a convergence criterion of 1.00 kcal mol^{-1} Å for 1000 iterations. The gradient type was kept as analytical with non bonded cut off value of 20.0 for electrostatic and 10.0 for VanDer Waal forces. These minimized ligand molecules were used for QSAR analysis.

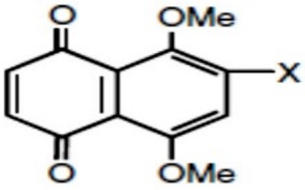
Molecular Alignment- Template based alignment method was used to align the naphthoquinone derivatives useful for studying shape variation with respect to the base structure selected for alignment which is useful for calculation of 3D descriptors. In this alignment method, a template structure is defined and used as a basis for alignment of a set of molecules. The reference molecule is required on which the other molecules of the align dataset get aligned based on the chosen template. The template structure was chosen based on the naphthoquinone scaffold (Figure 1a) and molecule number 1 was used as reference molecule.



Scaffold 1

Molecule No	X	Molecule No	X	Molecule No	X
1	H	4	COC_3H_7	7	COC_6H_{13}
2	COMe	5	COC_4H_9	8	$COCHMe_2$
3	COC_2H_5	6	COC_5H_{11}	9	$COCH_2CH_2CH=CH_2$

Table.1 Naphthoquinone derivatives for scaffold 1



Scaffold 2

Molecule No	X	Molecule No	X	Molecule No	X
10	CHO	14	COC_4H_9	18	COC_8H_{17}
11	$COCH_3$	15	COC_5H_{11}	19	OCC_9H_{19}
12	COC_2H_5	16	COC_6H_{13}	20	$COC_{10}H_{21}$
13	COC_3H_7	17	COC_7H_{15}		

Table.2 Naphthoquinone derivatives for scaffold 2

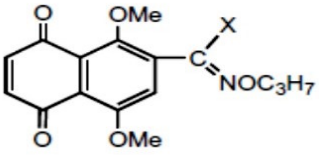
 Scaffold 3					
Molecule No	**X**	**Molecule No**	**X**	**Molecule No**	**X**
21	H	24	C_3H_7	28	C_7H_{15}
22	CH_3	25	C_4H_9	29	C_8H_{17}
23	C_2H_5	26	C_5H_{11}	30	C_9H_{19}
		27	C_6H_{13}		

Table.3 Naphthoquinone derivatives for scaffold 3

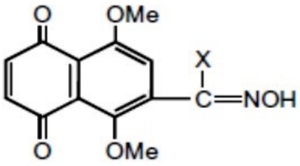
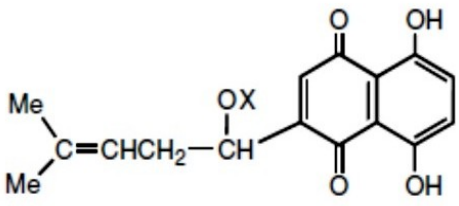
 Scaffold 4					
Molecule No	**X**	**Molecule No**	**X**	**Molecule No**	**X**
31	H	35	C_4H_9	39	C_8H_{17}
32	CH_3	36	C_5H_{11}	40	C_9H_{19}
33	C_2H_5	37	C_6H_{13}	41	$C_{10}H_{21}$
34	C_3H_7	38	C_7H_{15}	42	$C_{12}H_{25}$

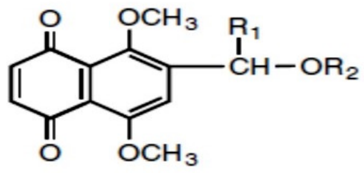
Table.4 Naphthoquinone derivatives for scaffold 4



Scaffold 5

Molecule No	X	Molecule No	X	Molecule No	X
43	Acetyl	48	Isobutanoyl	54	trans-3-Hexenoyl
44	Monochloroacetyl	49	n-Pentanoyl	55	2,4-Hexadienoyl
45	Trichloroacetyl	50	4-Pentenoyl	56	n-Heptanoyl
46	n-Propanoyl	51	trans-2-Pentenoyl	57	2,6-Heptadienoyl
47	n-Butanoyl	52	n-Hexanoyl	58	6-Heptenoyl
		53	trans-2-Hexenoyl		

Table.5 Naphthoquinone derivatives for scaffold 5



Scaffold 6

Molecule No	R1	R2	Molecule No	R1	R2	Molecule No	R1	R2
59	Me	H	70	C_2H_5	$CO(CH_2)_5CH_3$	80	C_4H_9	$CO(CH_2)_2CH_3$
60	Me	$COCH_3$	71	C_3H_7	H	81	C_4H_9	$CO(CH_2)_4CH_3$
61	Me	$COCH_2CH_3$	72	C_3H_7	$COCH_3$	82	C_4H_9	$CO(CH_2)_5CH_3$
62	Me	$CO(CH_2)_2CH_3$	73	C_3H_7	$COCH_2CH_3$	83	C_5H_{11}	$COCH_3$
63	Me	$CO(CH_2)_4CH_3$	74	C_3H_7	$CO(CH_2)_2CH_3$	84	C_5H_{11}	$COCH_2CH_3$
64	Me	$CO(CH_2)_5CH_3$	75	C_3H_7	$CO(CH_2)_4CH_3$	85	C_5H_{11}	$CO(CH_2)_2CH_3$
65	C_2H_5	H	76	C_3H_7	$CO(CH_2)_5CH_3$	86	C_5H_{11}	$CO(CH_2)_4CH_3$
66	C_2H_5	$COCH_3$	77	C_4H_9	H	87	C_5H_{11}	$CO(CH_2)_5CH_3$
67	C_2H_5	$COCH_2CH_3$	78	C_4H_9	$COCH_3$	88	C_7H_{15}	$COCH_3$
68	C_2H_5	$CO(CH_2)_2CH_3$	79	C_4H_9	$COCH_2CH_3$	89	C_6H_{13}	$COCH_2CH_3$
69	C_2H_5	$CO(CH_2)_4CH_3$				90	C_6H_{13}	$CO(CH_2)_2CH_3$

Table.6 Naphthoquinone derivatives for scaffold 6

Group based QSAR (GQSAR) - Here in, molecules were divided into six fragments based on the fragmentation rules derived in light of the specific molecular substitutions obtained from literature. In order to consider the environment of the neighboring fragment(s), the attachment point atoms were also included in the fragments. The scheme of molecular fragmentation is shown in Figure 3 and the final fragment template is indicated in Figure 4.

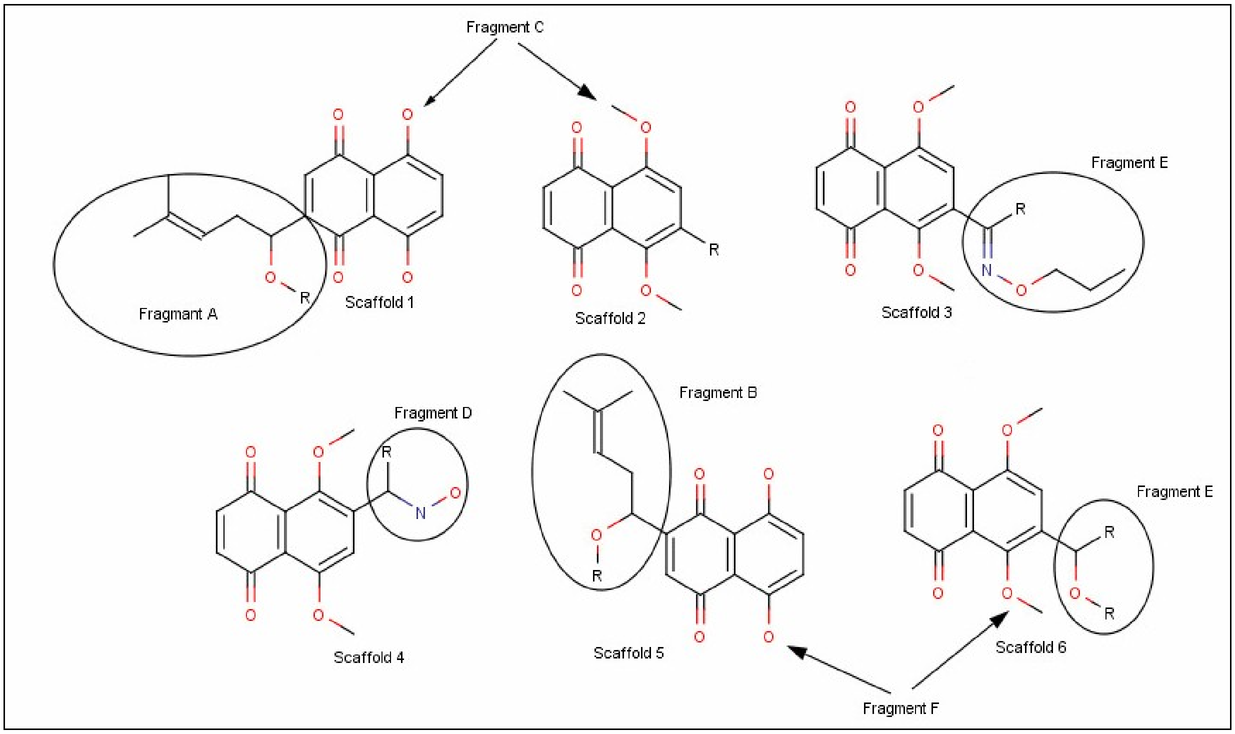


Figure.3 Fragmentation scheme for GQSAR

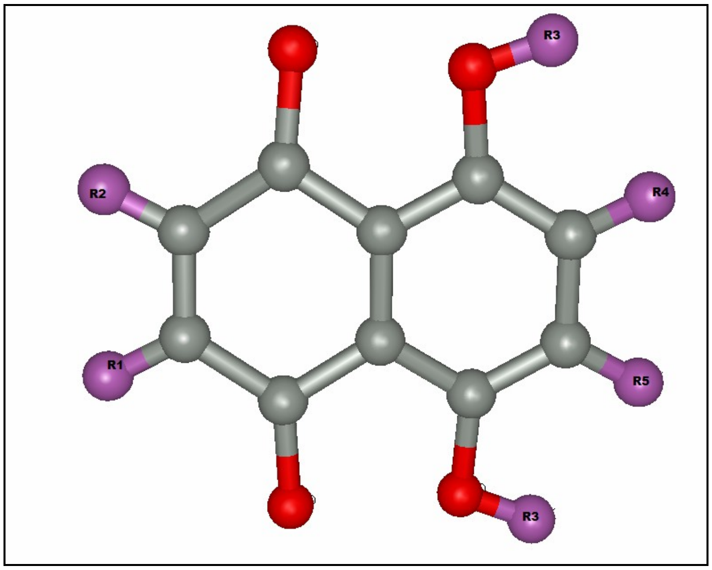


Figure.4 Template for fragmentation; purple points represent points of fragmentation

Fragment A- These are fragments obtained from 2-CH(OX)-($CH_2CH=CMe_2$)-5,8-dihydroxy-1,4-naphthoquinone derivatives.

Fragment B- These are fragments obtained from 2-CH(OX)-($CH_2CH=CMe_2$)-5,8-dihydroxy-1,4-naphthoquinone derivatives.

Fragment C- These are fragments obtained from the 6-X-5,8-dimethoxy-1,4-naphthoquinones.

Fragment D- These are fragments obtained from the 6-C(=NOH)X-5,8-dimethoxy-1,4-naphthoquinones.

Fragment E- These are fragments obtained from the 6-CH(R_1)(OR_2)-5,8-dimethoxy-1,4-naphthoquinones.

Fragment F- These are fragments obtained from the 6-C(=NOC_3H_7)X-5,8-dimethoxy-1,4-naphthoquinones.

Calculation of Molecular Descriptors

All the total two and three dimensional descriptors were calculated using VLifeMDS software for all of the 6 fragments [12]. These included various physicochemical (239), structural, topological, electro-topological, Baumann alignment independent topological descriptors (more than 700) [13] and atom type count descriptors (99). In addition various 3D descriptors such as electrostatic, hydrophobic and volume descriptors were also calculated. Preprocessing of the independent variables (i.e. descriptors) was done by removing the invariables (i.e. descriptor with a constant value for more than 95 percent molecules), which resulted in 1036 descriptors in the descriptor pool.

Creation of training and test set- Optimal training and test sets were generated using random selection algorithm keeping the selection percentage ratio as 80:20 for training and test set respectively. Seventy two compounds were used as training set and eighteen in the test set for the combined GQSAR analysis. The test set molecules were selected by considering the fact that this set of molecules represents a range of biological activity similar to that of the training set. Thus, the test set is the true representative of the training set. In order to assess the similarity of the distribution pattern of the molecules in the generated sets, statistical parameters (with respect to the biological activity) i.e. mean, maximum, minimum and standard deviation were calculated for the training and test sets.

Variable Selection Method- In order to select a subset of descriptors (variables) from the descriptor pool, a variable selection method known as stepwise forward backward selection was used [20,21].

The following techniques were used to develop the QSAR models

Multiple Regression Analysis (MLA) - Multiple regression is the standard method for multivariate data analysis. It is also called as ordinary least squares regression (OLS). This method of regression estimates the values of the regression coefficients by applying least squares curve fitting method. For getting reliable results, dataset having typically 5 times as many data points (molecules) as independent variables (descriptors) is required. The regression equation takes the form

$$Y = b_1*x_1 + b_2*x_2 + b_3*x_3 + c$$

Where Y is the dependent variable, the ‘b’s are regression coefficients for corresponding ‘x’s (independent variable), ‘c’ is a regression constant or intercept [18, 19].

Partial Least Square (PLS) Analysis- Partial least squares regression is an extension of the multiple linear regression model. In its simplest form, a linear model specifies the (linear) relationship between a dependent (response) variable Y, and a set of predictor variables, X's, so that

$$Y = b_0 + b_1X_1 + b_2X_2 + ... + b_pX_p$$

In this equation b_0 is the regression coefficient for the intercept and the bi values are the regression coefficients (for variables 1 through p) computed from the data. Partial least squares regression extends multiple linear regression without imposing the restrictions employed by discriminant analysis, principal components regression and canonical correlation. In partial least squares regression, prediction functions are represented by factors extracted from the Y'XX'Y matrix [14, 15].

Principal Component Analysis- it rotates the data into a new set of axes such that the first few axes reflect most of the variations within the data. By plotting the data on these axes, we can spot major underlying structures automatically. The value of each point, when rotated to a given axis, is called the principal component value. Principal Components Analysis selects a new set of axes for the data. These are selected in decreasing order of variance within the data. They are also perpendicular to each other. Hence the principal components are uncorrelated. Rather than forming a single model, as with MLR, a model can be formed using 1, 2 ... components and a decision can be made as to how many components are optimal [22-25].

k- Nearest Neighbour (k-NN) Analysis- The k-NN method was also used to develop a QSAR model using continuous variable i.e. using activity as pIC_{50} values. In this case, by using a developed k-NN QSAR model the activity of a molecule can be predicted using weighted average activity (Eq. (1)) of the k most similar molecules in the training set.

$$\hat{y}_i = \sum w_i y_i \qquad \text{………………… Eq 1}$$

Where y_i and $\hat{}y_i$ are the actual and predicted activity of the ith molecule respectively, and w_i are weights calculated using (Eq. (2)).

$$w_i = \frac{\exp(-d_j)}{\sum_{j=1}^{k} \exp(-d_j)} \qquad \text{……….. Eq 2}$$

The similarities were evaluated as the inverse of Euclidean distances (d_j) between molecules (Eq. (3)) using only the subset of descriptors corresponding to the model. Where, k is number of nearest neighbours in the model.

$$d_{i,j} = \left[\sum_{m=1}^{Vn} (X_{i,m} - X_{j,m}) \right]^{1/2} \qquad \text{. Eq 3}$$

Where, X is the matrix of selected descriptors (V_n) for the k-NN QSAR model [17].

Model Evaluation and Validation- This is done to test the internal stability and predictive ability of the QSAR models. Internal validation was carried out using leave-one-out (q^2, LOO) method. To calculate q^2, each molecule in the training set was sequentially removed, the model refit using same descriptors, and the biological activity of the removed molecule predicted using the refit model. The q^2 was calculated using Eq. (4).

$$q^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - y_{mean})^2} \qquad \text{….. Eq 4}$$

Where y_i, $\hat{}y_i$ are the actual and predicted activity of the i^{th} molecule in the training set, respectively, and y_{mean} is the average activity of all molecules in the training set. For external validation, activity of each molecule in the test set was predicted using the model generated from the training set. The pred_r^2 value is calculated as follows Eq. (5)

$$pred\_r^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - y_{mean})^2} \qquad \text{……. Eq 5}$$

Where y_i, $\hat{y}_i$ are the actual and predicted activity of the ith molecule in the 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. Thus the pred_r^2 value is indicative of the predictive power of the current model based on the external test set.

Developed quantitative models were evaluated using following statistical measures: n, number of observations (molecules); k, number of variables (descriptors); Number of components, number of optimum PLS components in the model; Number of nearest neighbours, number of k-nearest neighbour in the model; r^2, coefficient of determination; q^2, cross-validated r^2 (by leave one out); pred_r^2, r^2 for external test set; F-test, F-test value for statistical significance; SEE, standard error of estimate of the model; cv_SE, standard error of cross-validation and pred_SE, standard error of external test set prediction. The r^2 and q^2 values were used as deciding factors in selecting the optimal models.

Results and Discussion

Based on the information obtained from conventional 2D and 3D QSAR model descriptors, it is not exactly specified in which part of the molecule modifications are required so as to improve the activity, thus posing a hurdle in the complete structural interpretation. Therefore, in order to gain insight to the influential molecular part(s), in terms of their chemical information responsible for the variation in activity, GQSAR models involving fragment descriptors were developed. In addition to this, all the 2D and 3D descriptors were combined so as to obtain GQSAR models.

Using the molecular alignment technique all the molecules were aligned in their 3D space conformations (Figure 5). The molecules were fragmented in 6 parts depending upon the molecular substitutions and the scheme in Figure 3 keeping the alignment intact, so as not to disturb the space conformations of the molecules. Individual fragment based descriptors were calculated for all the 6 fragments.

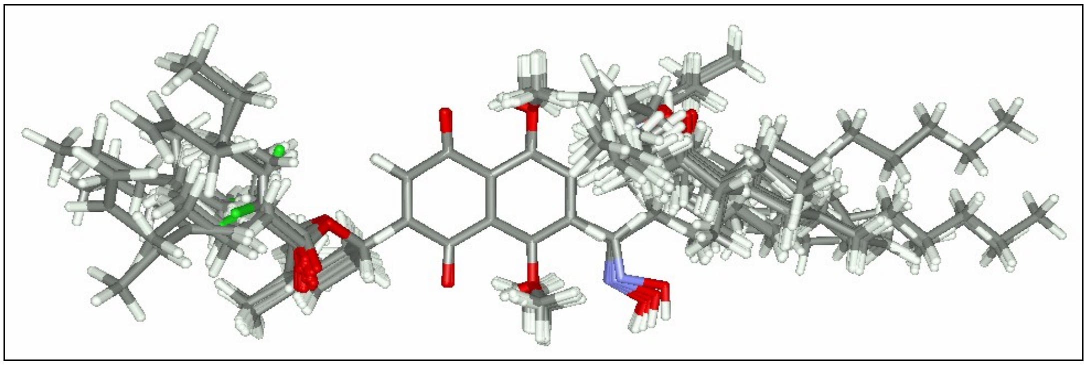


Figure 5- Molecular alignment of Naphthoquinone derivatives

The statistical parameters for assessing the distribution of activity in the training and test sets have been listed in Table 7. As can be seen from table, the minimum biological activity of test set is same as that the minimum activity of training set (not less) and the maximum activity of the test set is less than the maximum activity of the training set, this indicates that the test set is within the activity domain of the training set. The comparable standard deviation and the mean values (as shown in Table 7) of training and test sets show that there is a similar distribution of training and test set molecules with respect to the activity.

Parameters	Training set	Test set
Max	4.9400	4.8400
Min	3.200	3.200
Std. Dev	0.3170	0.3885
Mean	4.1628	4.3106

Table.7 Statistical parameters for assessing distribution of activity in training and test set

All the calculated descriptors remaining after preprocessing (1036) were subjected to step wise forward- backward variable selection coupled, separately, with MRA, PLS, PCA and k-NN methods for building 4 different QSAR models based on the same training set. This study led to various statistically significant 2D and 3D combined GQSAR models and their statistical parameters are reported in Table 8. Table 9 reports descriptors for each of the fragments with their regression coefficient and percentage contribution in each of the reported QSAR models.

Model Parameters	GQSAR MRA	GQSAR PLS	GQSAR PCR	GQSAR k-NN
Training set	72	72	72	72
Test set	18	18	18	18
R^2	0.7812	0.7214	0.5684	
Q^2	0.5622	0.5220	0.4704	0.5177
F-test	17.5509	23.6736	17.3820	
R^2_se	0.1707	0.1850	0.2267	
Q^2_SE	0.2415	0.2423	0.2511	0.2202
Pred_r^2	0.7575	0.4798	0.6845	0.4455
Pred_SE	0.1767	0.2572	0.2015	0.3107
Number of Descriptors k	12	9	6	3
Number of components/nearest neighbor	72	7	5	72
Degree of freedom	59	64	66	68

Table.8 Statistical parameters of various GQSAR models

The equations explain 78 % (r^2 = 0.78) and 72% (r^2=0.72) of the total variance in the training set for the MRA and PLS models respectively. It also has an internal (q^2) predictive ability of ~56 % and ~52% and external (pred_r^2) predictive ability of 75% and 47% respectively. The F-test = 17.68 and 23.67 for MRA and PLS models respectively shows the statistical significance of the model which means that probability of failure of the model is very less. For the PCA model the r^2 decreased to 0.56 indicating 56% of the total variance in the training set. Also the q^2 and pred_r^2 values indicate 47% and 68% of predictive ability for the model. The F-test = 17.38 shows the statistical significance of the model which means that probability of failure of the model is very less.

Descriptor	**MLR coefficient**	**Percent contribution**	**PLS coefficient**	**Percent contribution**	**PCR coefficient**
R1-SssCH2E-index	-0.1329	-7.20			-0.1004
R2-MMFF_6	-0.3565	-6.11	-0.3247	-13.78	
R3-Quadrupole3	0.0373	2.80			
R2-HosoyaIndex	-0.006	-3.91	-0.0011	-17.57	-0.0005
R2-G_2_T_5	0.0996	4.96	0.0824	10.13	
R1-H-DonorCount	-0.7940	-3.46	-0.8905	-9.25	0.5970
R2-G_C_C_7	-0.2325	-15.48			
R1-MomInertiaY	-0.0002	-8.76			
R1-T_2_O_6	0.3153	7.27			
R2-G_T_T_6	0.1211	13.90			
R1-G_T_O_4	0.2251	13.42			
R1-G_C_C_6	-0.1529	-12.73			
R1-HosoyaIndex			-0.0018	-28.99	
R1-Quadrupole2			-0.0788	-7.97	
R1-SdCH2E-index			0.0798	5.74	
R2-G_T_O_7			0.1091	6.57	
R5-Quadrupole1					-5.4357
R3-Quadrupole2					-0.0538
R1-XKMostHydrophilic					-1.6114

Table.9 Descriptors from MRA, PLS and PCR models with their coefficients

The contributions of the individual descriptors for both MRA and PLS are reported in Figure 6. Figure 7 shows the comparison of percentage contribution of descriptors common between various QSAR models such as MRA and PLS and MRA and PCA. From the figure it can be seen that common descriptors are R2_MMFF6, R2_Hosoyaindex and R2_G_2_T_5 for fragment B and R1_Hdonorcount for fragment A. It was seen that all the descriptor contributions for both the MRA and PLS are relatively same. Also all the fragments except R2_G_2_T_5 contribute negatively for the activity. For the MRA and PCA the common descriptors are R1_SssCH2Eindex and R1_Hdonorcount for fragment A and R2_Hosoyaindex for Fragment B. Here also the comparative contributions for both the analysis are relatively similar. The descriptors R2_Hosoyaindex and R1_Hdonorcount are common for all the 3 analysis. Table 10 reports the list of important fragment specific descriptors found in various QSAR models along with their descriptor category and definition and Figure 8 shows summary of all the important descriptors.

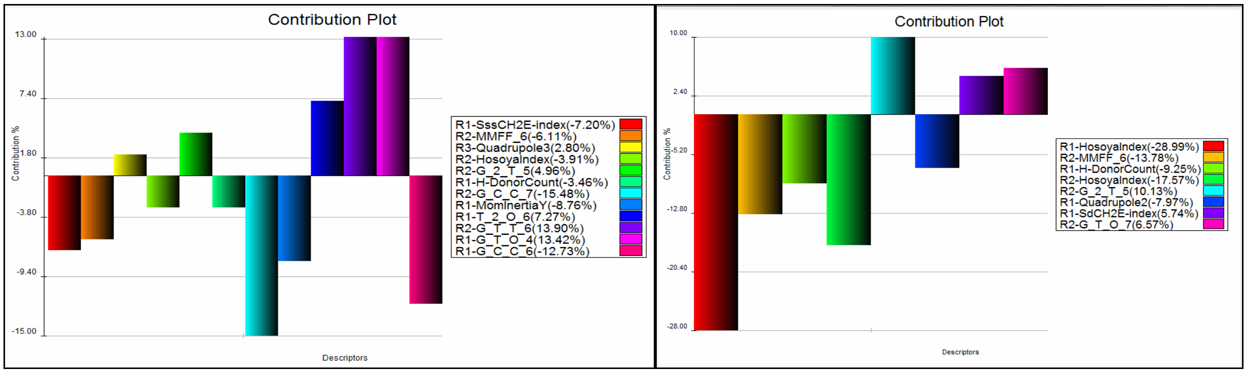


Figure.6 contribution plot of descriptors (left) MRA model, (right) PLA model

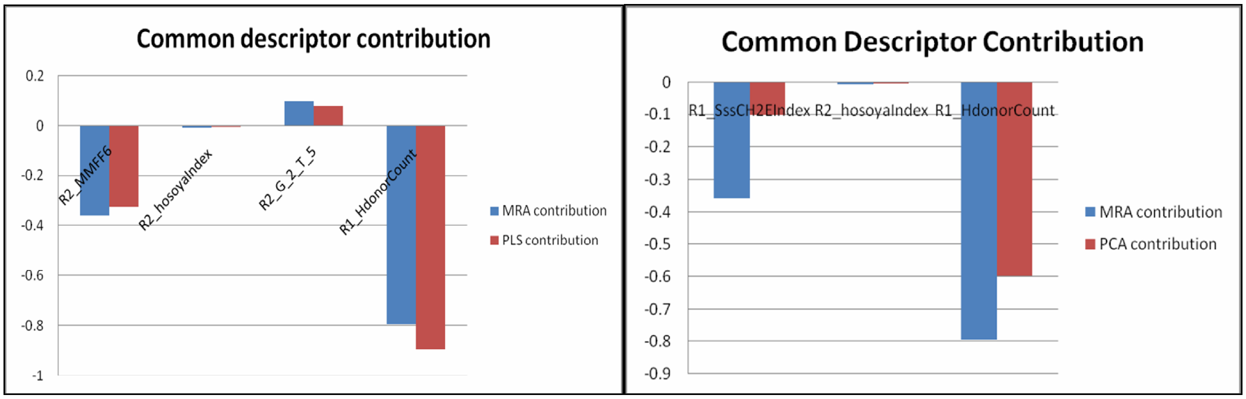


Figure.7 Plot of contribution of descriptors common to (left) MRA and PLS model, (right) MRA and PCA model

Table.10 List of descriptors along with their category and definition

Descriptors	Category	Definition
Fragment A		
R1-SssCH2E-index	Estate contributions	Electrotopological state indices for number of –CH2 group connected with two single bonds.
R1-H-DonorCount	Physicochemical	Number of hydrogen bond donor atoms
R1-MomInertiaY	Distance based Topological	This descriptor signifies moment of interia at Y-axis
R1-T_2_O_6	Alignment independent topological	Count of number of double bounded atoms separated from Oxygen atom by 6 bonds.
R1-G_T_O_4	Geometrical Topological	Count of number of topological atoms separated from oxygen atom by 4 bonds
R1-G_C_C_6	Geometrical Topological	Count of number of carbon atoms separated from each other by 6 bonds
R1-HosoyaIndex	Distance based Topological	signifies the topological index or Z index of a graph is the total number of matching in it plus 1 ("plus 1" accounts for the number of matchings with 0 edges)
R1-Quadrupole2	Dipole Moment	Signifies magnitude of second tensor of quadrupole moments.
R1-SdCH2E-index	Estate contributions	Electrotopological state indices for number of –CH2 group connected with one double bond.
R1-XKMostHydrophilic	Hydrophobicity XlogpK	Most hydrophilic value on the vdW surface
Fragment B		
R2-MMFF_6	Merck molecular force field (MMFF) atom type	Count of beta carbon in 5-membered hetero-aromatic ring
R2-HosoyaIndex	Distance based Topological	signifies the topological index or Z index of a graph is the total number of matching in it plus 1 ("plus 1" accounts for the number of matchings with 0 edges)
R2-G_2_T_5	Geometrical Topological	
R2-G_C_C_7	Geometrical Topological	Count of number of carbon atoms separated from each other by 7 bonds
R2-G_T_T_6	Geometrical Topological	
R2-G_T_O_7	Geometrical Topological	Count of number of topological atoms separated from oxygen atom by 7 bonds
Fragment C		
R3-Quadrupole3	Dipole Moment	signifies magnitude of third tensor of quadrupole moments
R3-Quadrupole2	Dipole Moment	Signifies magnitude of second tensor of quadrupole moments.
Fragment E		
R5-Quadrupole1	Dipole Moment	Signifies magnitude of first tensor of quadrupole moments.

It can be seen from Figure 6 and Table 10, that substitution on fragment A and B are most influencing with highest percentage contribution in all the GQSAR models. This is also supported by the fact that the largest amount of variation in the chemical substituents is contained in fragment A followed by fragment B than fragments E and D. The activity variation is explained in terms of the Baumann's alignment independent topological descriptors, geometrical topological descriptors, dipole moment descriptors and other basic descriptors. Also, descriptors influencing the activity in favourable and unfavourable ways were found to be near 45 percent and 55 percent, respectively.

This information suggests that there is almost equal opportunity to optimize both the favourable and unfavourable descriptors in the design of new molecules. It is found that most of the contributing descriptors from all the models are from Fragment A. The important descriptors contributing towards Fragment A are R1-SssCH2E-index, R1-H-DonorCount, R1-MomInertiaY, R1-T_2_O_6, R1-G_T_O_4, R1-G_C_C_6, R1-HosoyaIndex, R1-Quadrupole2, R1-SdCH2E-index and R1-XKMostHydrophilic. Out of these R1-T_2_O_6, R1-G_T_O_4 and R1-SdCH2E-index contribute positively and others contribute negatively towards the inhibitory activity of Naphthoquinones against Topo I.

R1-SdCH2E-index indicates the importance of substituted double bonded carbon atom (CH2=) to increase the activity. In the same way the descriptors R1-T_2_O_6 and R1-G_T_O_4 are directly proportional to the activity as indicated in the Multiple Regression GQSAR model. The descriptor R1-T_2_O_6 shows the importance of double bonded C atoms separated from Oxygen atom by 6 bonds at fragment A to be detrimental to the inhibitory activity of the Naphthoquinone derivatives. The geometrical topological descriptor R1-G_T_O_4 contributes most positively towards the activity with a contribution of 13.42 %.

It represents the geometrical topological index value for C atom separated from Oxygen by 4 bonds which is important for increasing the inhibitory activity. The remaining molecular descriptors are inversely proportional to the inhibitory activity and the most negatively contributing descriptor is R1-G_C_C_6 with -12.73% and R1-HosoyaIndex with -28.99%. The next important fragment contributing towards the overall inhibitory activity is fragment B. The descriptors important are R2-MMFF_6, R2-HosoyaIndex, R2-G_2_T_5, R2-G_C_C_7, R2-G_T_T_6 and R2-G_T_O_7.

Out of these the positively contributing descriptors are the geometrical topological indices such as R2-G_2_T_5, R2-G_T_T_6 and R2-G_T_O_7. All the remaining descriptors are inversely proportional to inhibitory activity. The most influencial geometrical topological descriptor is R2-G_T_T_6 with 13.90% contribution followed by R2-G_2_T_5 with 10.13% as indicated in the Partial Least Square Analysis model. The most negatively contributing descriptor is R2-HosoyaIndex with –17.57% contribution followed by -15.48% contribution by R2-G_C_C_7.

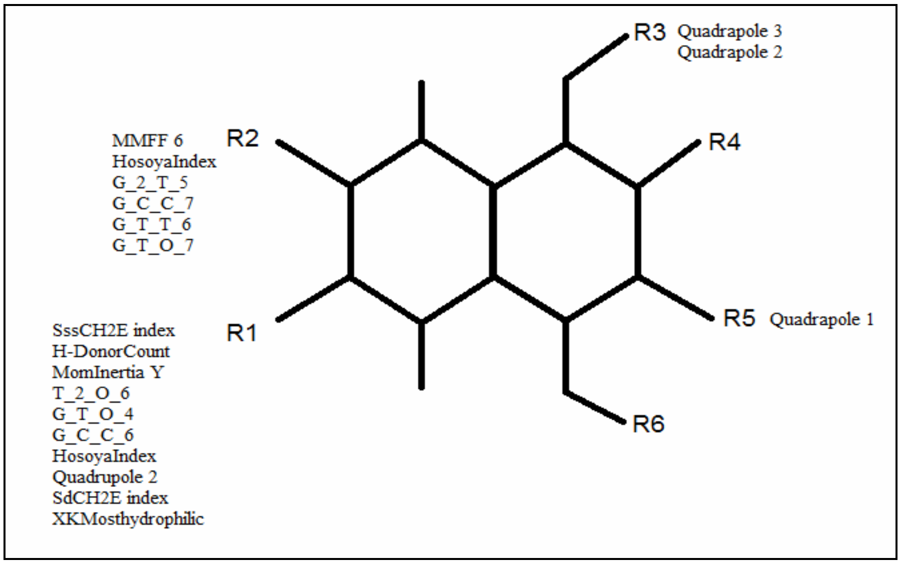


Figure 8 - Summary of specific descriptors for different fragments

Figure.9-12 show the observed versus predicted biological activity plot of training and test set molecules by all GQSAR models. The plot of observed vs. predicted activity provides an idea about how well the model was trained and how well it predicts the activity of the external test set. From the plot it can be seen that model is able to predict the activity of training set quite well (all points are close to regression line) as well as external test set up to ~60% (only 1 point is relatively apart from the regression line) in the PCA model providing confidence in predictive ability of the model.

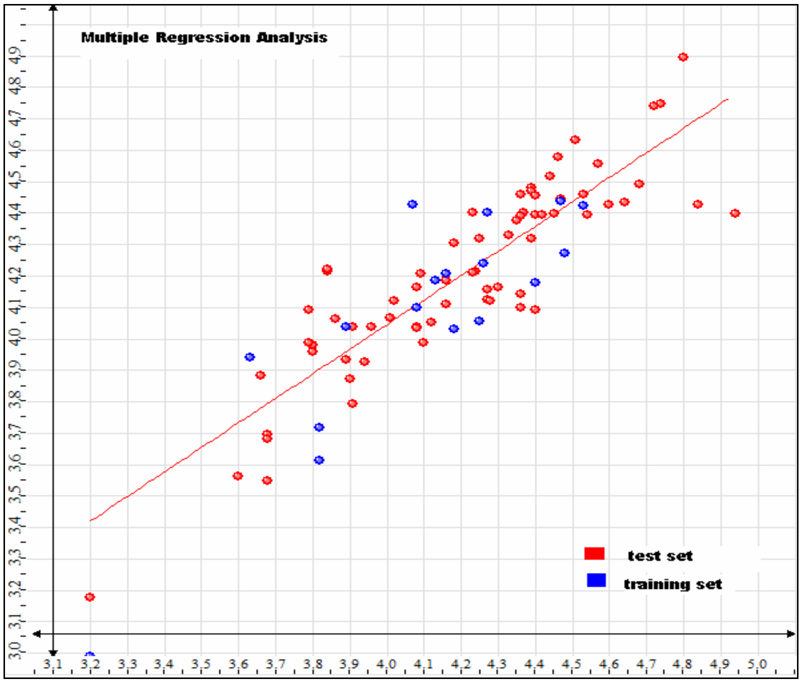


Figure.9 Plot of observed versus predicted pIC_{50} values obtained from MRA GQSAR model

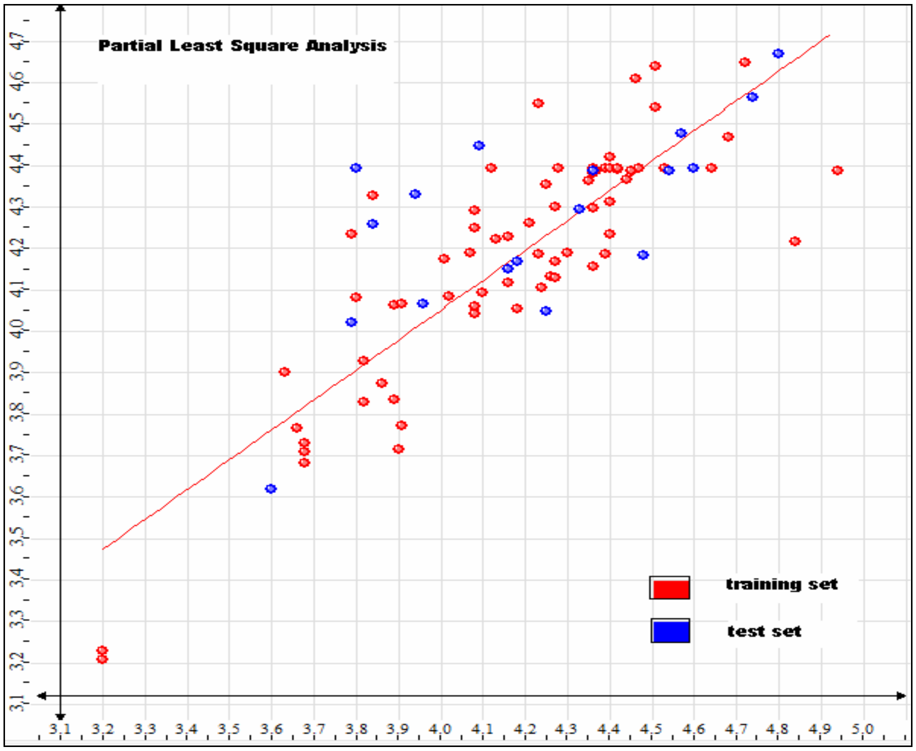


Figure.10 Plot of observed versus predicted pIC_{50} values obtained from PLS GQSAR model

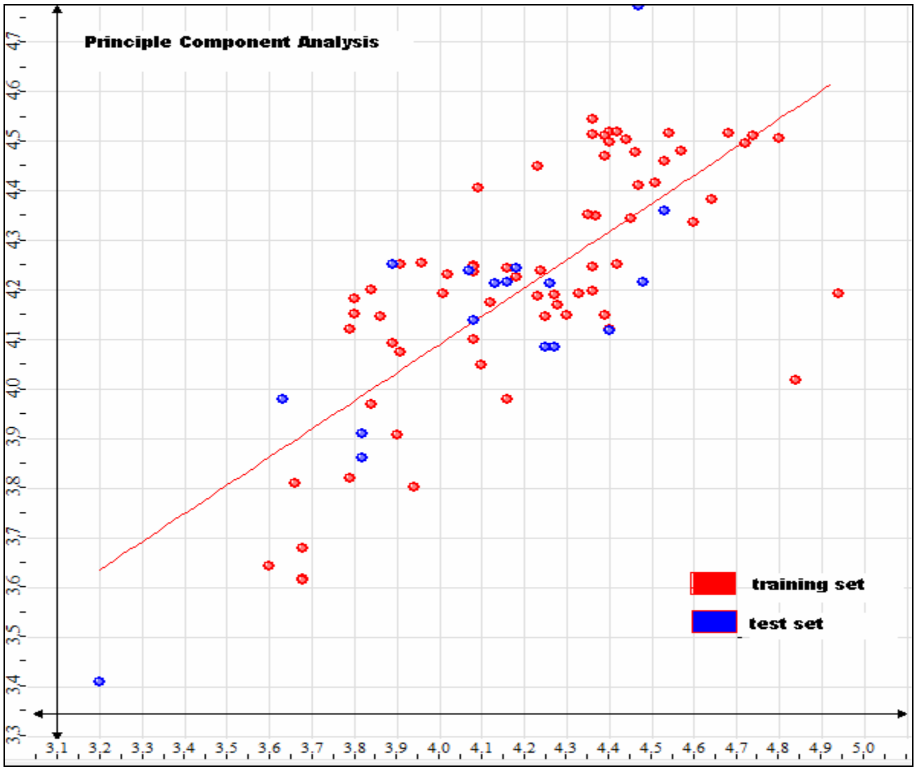


Figure.11 Plot of observed versus predicted pIC_{50} values obtained from PCA GQSAR model

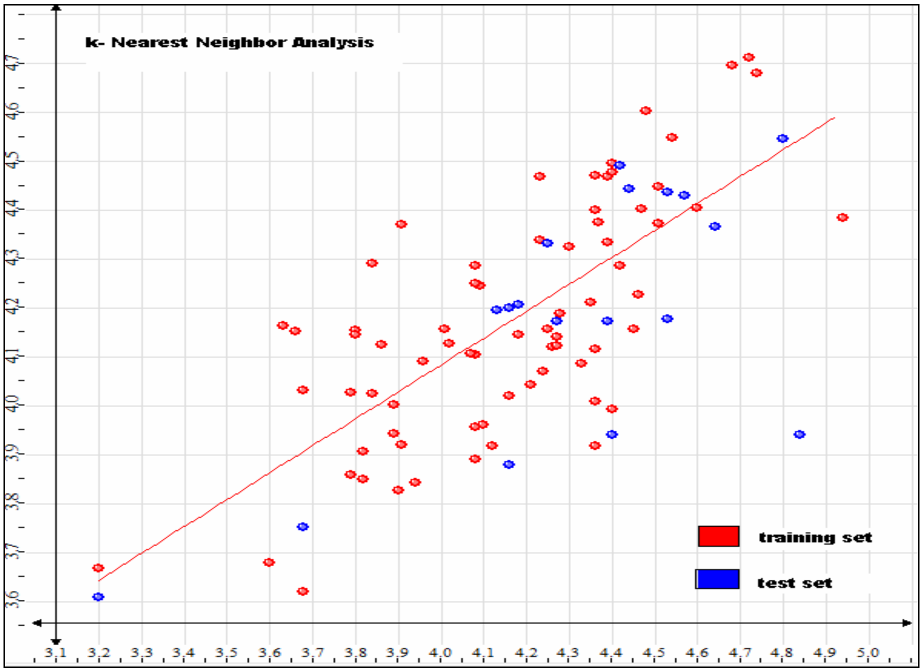


Figure.12 Plot of observed versus predicted pIC_{50} values obtained from k-NN GQSAR model

Table 8 shows that the best 2D, 3D combined GQSAR model was derived from both the multiple regression analysis and partial least square analysis and it was found to have improved statistical parameters as compared to GQSAR PCR model. Hence, we have also developed GQSAR k-NN model by subjecting all the calculated fragment descriptors to the step wise forward backward selection coupled with k-NN method, to capture nonlinearity in terms of individual fragment descriptors. This study has resulted in a k-NN GQSAR model which was found to be comparable to above reported GQSAR MRA and PLS model but has lower statistical significance (with respect to pred_r^2) as compared to GQSAR models. The descriptors that were found to be important in the k-NN GQSAR model are: R6-MomInertiaY, R2-Quadrupole2, and R1-HydrogensCount.

Descriptor	Range
R6-MomInertiaY	942.0660 944.8230
R2-Quadrupole2	-9.7037 -4.4215
R1-HydrogensCount	2.0000 2.0000

Table.11 k-NN model descritptors with their minimum and maximum ranges

An advantage of the k-NN method is that it can provide ranges (minimum and maximum, derived from the k nearest neighbours of the most active molecule) for each fragment descriptor as reported in Table 11. These ranges can be used as a reference when searching for similar fragments in a fragment database during the design of new molecules. Thus, unlike traditional QSAR models, the developed combined 2D and 3D GQSAR models provide information about the important substitution site(s) along with their chemical nature and their interactions which could prove useful for designing of new molecules. Figure 13 helps us to identify the important features required at various positions so as to obtain a better lead molecule showing anti topoisomerase 1 activity. These features can be incorporated so as to design potential lead molecules.

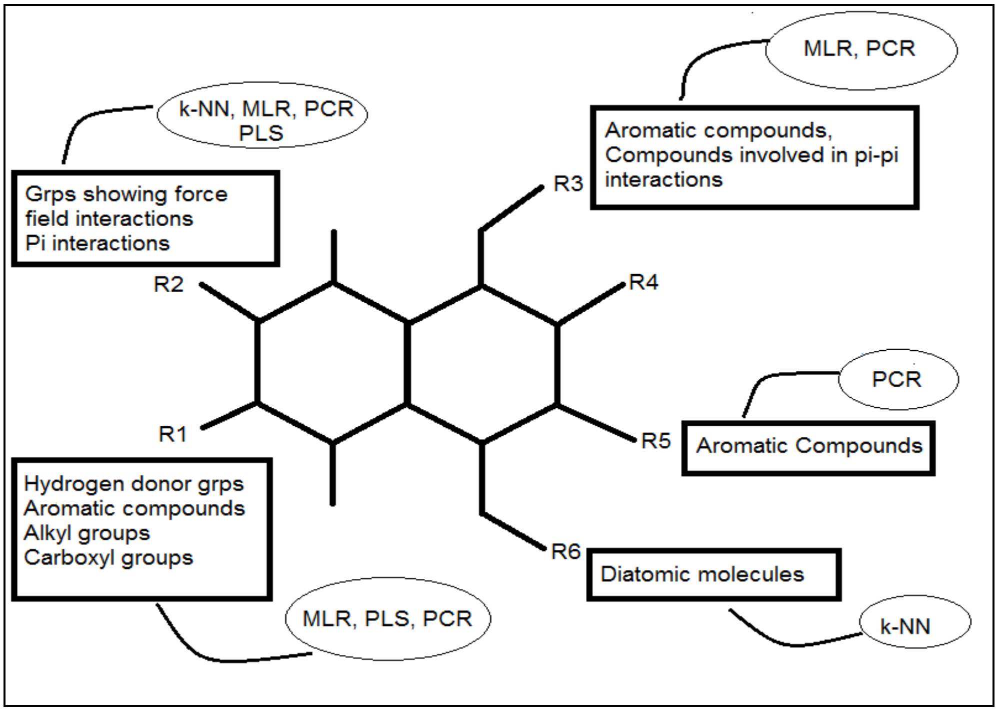

MLR, PCR
k-NN, MLR, PCR PLS
Aromatic compounds, Compounds involved in pi-pi interactions
R3
Grps showing force field interactions Pi interactions
R2
R4
PCR
R1
R5
Aromatic Compounds
Hydrogen donor grps Aromatic compounds Alkyl groups Carboxyl groups
R6
Diatomic molecules
MLR, PLS, PCR
k-NN

Figure.13 Schematic representation of different features at various position favouring anti topoisomerase activity

Conclusion

The present study unveils key structural requirements for Topoisomerase I inhibition utilizing various GQSAR methods. A wide variety of structurally diverse Topoisomerase I inhibitors (naphthoquinone derivatives) collected from various literature reports were used in this study. The GQSAR analyses revealed the major importance of Baumann’s alignment independent topological descriptors and geometrical topological indices along with other descriptors such as number of hydrogen bond donors, number of hydrogens, moment of inertia, Hosoyaindex in determining Topoisomerase I inhibition activity. The study reveals that any substitution on fragment A and fragment B will enhance the activity of the Naphthoquinone derivative against Topo I. Thus a combined Naphthoquinone derivative having both these fragments can be of importance for the inhibitory activity on Topo I.

Acknowledgements

We express our gratitude to European Union Erasmus Mundus Mobility for Life for granting us Fellowship under Host University, University of Rome ‘Tor Vergata’, Italy. We express special Thanks to Prof. Alessandro Desideri and his team for providing us infrastructure and guidance.

References

1. Babula, P; Adam, V; Havel, L; Kizek, R (2007), Ceska a Slovenska farmacie: casopis Ceske farmaceuticke spolecnosti a Slovenske farmaceuticke spolecnosti 56 (3): 114–20.
2. Ting, C.-Y.; Hsu, C.-T.; Hsu, H.-T.; Su, J.-S.; Chen,T.-Y.; Tarn, W.-Y.; Kuo, Y.-H.; Whang-Peng, J.; Liu, L. F.; Hwang (2003), J. Biochem. Pharmacol., 66, 1981.
3. Song, G. Y.; Kim, Y.; You, Y.-J.; Cho, H.; Kim, S.-H.; Sok, D.-E.; Ahn, B.-Z. (2000) Arch. Pharm. Pharm. Med. Chem., 333, 87.
4. Chae, G.-H.; Song, G.-Y.; Kim, Y.; Cho, H.; Sok, D.-E.; Ahn, B.-Z.(1999) Arch. Pharm. Res., 22, 507.
5. Song, G.-Y.; Kim, Y.; Zheng, X.-G.; You, Y.-J.; Cho, H.; Chung, J.-H.; Sok, D.-E.; Ahn, B.-Z (2000). Eur. J. Med. Chem., 35, 291.
6. Song, G.-Y.; Zheng, X.-G.; Kim, Y.; You, Y.-J.; Sok, D.-E.; Ahn, B.-Z (1999). Bioorg. Med. Chem. Lett., 9, 2407.
7. Kim, Y.; You, Y.-J.; Ahn, B.-Z. (2001) Arch. Pharm. Pharm. Med. Chem., 334, 318.
8. Fesen, M. R.; Kohn, K. W.; Leteurtre, F.; Pommier, Y. (1993) Proc. Natl. Acad. Sci. USA, 90, 2399.
9. S. Ajmani, K. Jadhav, S.A. Kulkarni, (2009) QSAR Comb. Sci. 28 36–41.
10. Ahn, B.-Z.; Sok, D.-E.(1996) Curr.Pharm.Des., 2, 247.
11. Ahn, B.-Z.; Baik, K.-U.; Kweon, G.-R.; Lim, K.; Hwang, B.-D (1995). J.Med.Chem., 38, 1044.
12. VLifeMDS, Version 3.5, VLife Sciences Technologies Pvt. Ltd., Pune, India, (2008).
13. K. Baumann, (2002) J. Chem. Inf. Comput. Sci. 42 26–35.
14. S. Wold, QSAR-Chemometric Methods in Molecular Design, vol. 2, Wiley–VCH, Weinheim, Germany, 1995, pp. 195–218.
15. S. Wold, A. Ruhe, H. Wold, W.J. Dunn, (1984) , SIAM J. Sci. Stat. Comp. 5 735–743.
16. S. Ajmani, K. Jadhav, S.A. Kulkarni, (2009) , QSAR Comb. Sci. 28 36–41.
17. M.A. Sharaf, D.L. Illman, B.R. Kowalski, (1986) Wiley, New York,.
18. M. H. Kutner, C. J. Nachtsheim, and J. Neter (2004), "Applied Linear Regression Models", 4th ed., McGraw-Hill/Irwin, Boston.
19. N. Ravishankar and D. K. Dey (2002), Chapman and Hall/CRC, Boca Raton.
20. Foster, Dean P. and Edward I. George (1994), *Annals of Statistics* Volume 22, Number 4 1947-1975. doi:10.1214/aos/1176325766.
21. Wilkinson, L. and Dallal, G.E. (1981), *Technometrics. 23*. 377-380
22. Pearson, K. (1901). *Philosophical Magazine* **2**(6): 559–572.
23. Shaw PJA (2003) *Multivariate statistics for the Environmental Sciences*, Hodder-Arnold. ISBN 0-3408-0763-6
24. A. A. Miranda, Y. A. Le Borgne, and G. Bontempi.(2008) New Routes from Minimal Approximation Error to Principal Components, Volume 27, Number 3 / June, , Neural Processing Letters, Springer
25. Fukunaga, Keinosuke (1990). *Introduction to Statistical Pattern Recognition*. Elsevier. ISBN 0122698517.