Canadian Journal of Chemistry
Revue canadienne de chimie

**Exploring structural requirements of unconventional
Knoevenagel-type indole derivatives as anticancer agents
through comparative QSAR modeling approaches**

Journal:	<i>Canadian Journal of Chemistry</i>
Manuscript ID	cjc-2016-0050.R1
Manuscript Type:	Article
Date Submitted by the Author:	18-Apr-2016
Complete List of Authors:	AMIN, SK. ABDUL; Dr. Harisingh Gour University Adhikari, Nilanjan ; Jadavpur University Jha, Tarun; Jadavpur University gayen, shovanlal; Dr. Harisingh Gour University
Keyword:	Knoevenagel-type indoles, cytotoxicity, 2D-QSAR, kNN-MFA, CoMSIA

SCHOLARONE™
Manuscripts

**Exploring structural requirements of unconventional Knoevenagel-type indole derivatives
as anticancer agents through comparative QSAR modeling approaches**

Sk. Abdul Amin, Nilanjan Adhikari, Tarun Jha and Shovanlal Gayen

Draft

S. A. Amin and S. Gayen. Laboratory of Drug Design and Discovery, Department of Pharmaceutical Sciences, Dr. Harisingh Gour University (A Central University), Sagar 470003, (MP), India.

N. Adhikari and T. Jha. Natural Science Laboratory, Division of Medicinal and Pharmaceutical Chemistry, Department of Pharmaceutical Technology, P.O. Box 17020, Jadavpur University, Kolkata 700032, (WB), India.

Corresponding author: Tarun Jha (e-mail: tjjupharm@yahoo.com) and Shovanlal Gayen (e-mail: shovanlal.gayen@gmail.com)

Abstract

Indole ring system is considered as a versatile scaffold in the pharmaceutical field. In this article, comparative QSAR modeling (2D-QSAR, 3D-QSAR; kNN-MFA and CoMSIA) was performed on some Knoevenagel-type cytotoxic indole derivatives to understand the structural requirements for the cytotoxic property of these compounds. The 2D-QSAR model was statistically significant and imparted high predictive ability ($n_{Train} = 30$; $R = 0.917$; $R^2_A = 0.801$; $cR^2_p = 0.757$; $Q^2 = 0.722$; $n_{Test} = 9$; $R^2_{pred} = 0.799$). Statistically significant 3D-QSAR kNN-MFA model (both with stepwise forward and simulated annealing model selection method) as well as 3D-QSAR CoMSIA model was developed to identify the key chemical features associated with enhancing the cytotoxic activities of these indoles. The results suggest that the presence of bulky group in R position can cause better cytotoxic activities. Consequently, substitution with cyano group at X position and cyano/ester/keto/sulphonyl features at Y position is favourable for the cytotoxicity. However, hydrophobic features in R' region are unfavourable for the biological activity. The chemical and structural features identified from the study may provide important avenues to modulate the structure of these indoles to a desirable biological end point.

Keywords: Knoevenagel-type indoles; cytotoxicity; 2D-QSAR; kNN-MFA; CoMSIA.

Graphical abstract

Introduction

Heterocyclic molecules represent essential substructures or backbone for the synthesis of many therapeutically important synthetic and natural compounds.¹ Moreover, heterocyclic ring is present in many drugs. Indole scaffold one of the most important heterocyclic structural subunits has aroused great attention because of its significance in therapeutic importance including anti-inflammatory, antifungal, antibacterial, antiviral, analgesic, anticonvulsant, antimalarial, anti-cancer activity etc.¹ Chemically indole is known as benzopyrrole in which the aromatic benzene nucleus is fused with heterocyclic pyrrole ring through the 2- and 3-positions of the pyrrole scaffold.² When the terminal exocyclic double bond carbon of 3-vinyl indole contains two electronegative groups, this type of indole is designated as Knoevenagel-type indoles.³ Knoevenagel-type indoles are found to be promising scaffold as this ring system may serve as several promising biological potential against many diseases.³ However, no QSAR study has been published up to date, which gives complementary insight into the structural and physicochemical interactions of these Knoevenagel-type indoles that require for chemical and biological processes.

In this article, multiple validated comparative molecular modeling approaches (such as 2D-QSAR, kNN-MFA QSAR and CoMSIA QSAR) were performed on unconventional Knoevenagel-type indoles.³ The findings may enlighten the coordinated application of different QSAR models yielding significant and complementary insights into the structural and physicochemical requirement for the cytotoxicity and may provide an efficient and general way to predict the cytotoxic potency of the Knoevenagel-type indoles.

Material and Methods

Dataset

A dataset comprising 39 unconventional Knoevenagel-type indole derivatives having MT-4 cell-based cytotoxic property³ (shown in Table 1) was considered for the chemometric modelling study as a part of our rational drug design program.⁴⁻¹¹ The reported EC₅₀ values (μM) of these compounds were subsequently converted to the corresponding negative logarithmic scale [pEC_{50}] for QSAR study.¹²⁻¹⁷ The chemometric modeling studies was performed by using VLifeMDS 4.4 software¹⁸ and SYBYL-X 2.0 software¹⁹ on a HP computer with Window 7 operating system.

[Table 1 may be placed here]

Statistical methods and 2D-QSAR model development

The 2D structures of these indole derivatives were drawn (shown in Figure 1) and subsequently converted to their corresponding 3D structures.²⁰ Energy of these molecules was minimized by batch energy minimization method with RMS gradient (convergence criterion) of 0.01 kcal/mol using a MMFF94s (Merck molecular force field) force field and charge, maximum number of cycles were 1,000, medium's dielectric constant of 1. The energy-minimized geometry was considered for calculation of descriptors such as physio-chemical (like electronic, thermodynamic, spatial) as well as structural descriptors, (e.g., retention index (chi), chain path count, path count, atomic valence connectivity index (chiV), element count, path cluster, semi-empirical, estate number, partition coefficient, molecular weight, molecular refractivity, and topological index) and Baumann alignment-independent (AI) descriptors. Finally the dataset was

divided into test and training set by *k*-means clustering analysis (*k*-MCA) ensuring structural and biological activity variation among both the sets.²¹ Stepwise multiple linear regression (S-MLR) methods¹²⁻¹⁷ used as a statistical tool for building multistep 2D-QSAR models step by step for these indole-based derivatives. During model development, the biological activity parameter was considered as the dependent variable and the physicochemical parameters were considered as the independent variables. Several internal and external validation parameters were considered to judge the predictability, significance and robustness of this models.²²

[Figure 1 may be placed here]

Validation of developed QSAR models

Leave-one-out (LOO) cross validation method²³ was applied as the internal validation tools to validate the developed 2D-QSAR models. The predictive powers of all the developed QSAR models were justified by this internal cross-validation technique. As per the LOO- cross validation technique, each molecule was deleted from the training set once and the prediction of the activity of the deleted molecule was done simultaneously by using the model developed from the remaining training set molecules.⁴⁻⁷ The internal predictability parameter (Q^2) of the QSAR model was calculated by eq. (1) shown below:

$$Q^2 = 1 - \frac{\sum(y_i - y_{calc})^2}{\sum(y_i - y_{mean})^2} \quad \text{Eq. (1)}$$

Where y_i and y_{calc} 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 these molecules in the training set.

Further, the external predictive ability of the developed QSAR models was checked by the R_{pred}^2 parameter as per eq. (2):

$$R_{pred}^2 = 1 - \frac{\sum(y_i - y_{pred})^2}{\sum(y_i - y_{mean})^2} \quad \text{Eq. (2)}$$

Where y_i and y_{pred} are the actual and predicted activity of the ' i 'th molecule in the test set, respectively, and y_{mean} is the average activity of all the molecules present in the training set.

The rm^2 matrices ($rm^2_{(LOO)}$, $rm^2_{(test)}$ and $rm^2_{(overall)}$) were calculated to check the closeness between the values of the predicted and the corresponding observed values of the training set, test set and overall set respectively. The robustness of the 2D-QSAR equation was validated by calculating the Y-randomization cR^2_p values. Apart from that, other statistical parameters were also considered to judge as the internal and external validation criteria for judging the significance of the model.²³⁻²⁵

3D-QSAR modelling

kNN-MFA 3D-QSAR technique

Alignment of molecules

These Knoevenagel-type indole molecules were subjected to molecular alignment by VLife MDS software¹⁸ to explore optimal alignment between the molecular structures requirement for ligand-receptor interactions.²⁶ Molecules were aligned by template-based method. The most potent bioactive compound 35 was chosen as a reference molecule on which other molecules in the data set were aligned to obtain conformational flexibility in 3D space.

Calculation of field descriptors

The aligned biologically active conformations of indole-based derivatives were employed to calculate the steric, electrostatic and hydrophobic type molecular fields. For the calculation of molecular fields, three-dimensional rectangular grids were generated around the molecule and the interaction energy between the molecule and probe group placed at each grid point were calculated. The sp^3 carbon probe atom having charge +1.0 with 10.0 kcal/mole electrostatic and 30.0 kcal/mole steric cut-off was utilised to generate fields. The dielectric constant was set to 1.0 and steric, electrostatic fields at each grid point were computed considering Gasteiger-Marsili charge. Each aligned bioactive conformation of the investigated compound was taken in turn and the molecular fields around it were calculated. A huge number of descriptors were calculated for all the compounds. Finally, the kNN-MFA 3D-QSAR models were generated after exclusion of all the invariable descriptors having zero values or same values, as these descriptors were unable to contribute to the QSAR models.²⁷

k-Nearest neighbor molecular field analysis (kNN-MFA)

The kNN methodology simply based on the distance learning approach in which an unknown member is classified according to the majority of its k-nearest neighbours (kNN) in training set. The nearness is measured by an appropriate distance metric (e.g., a molecular similarity measure computed using field interactions of molecular structures). In kNN-MFA technique, several models were generated on training sets and subsequent validated on test sets. Once training and test sets are originated, kNN technique is used to generate descriptors over the grid.²⁷ The kNN-MFA models were generated by stepwise forward (SWF) as well as by simulated annealing (SA) methods.²⁷

3D-QSAR comparative molecular similarity analysis (CoMSIA) technique

Comparative molecular similarity analysis (CoMSIA) technique deals with the calculation of the similarity indices at the intersections of a surrounding lattice.²⁶ Important assumption is taken in the CoMSIA study that the changing of the binding affinities of the structurally aligned molecules are related to the changes in molecular properties represented by different fields (steric, electrostatic hydrophobic, hydrogen bond donor, and hydrogen bond acceptor). In CoMSIA, a probe sp³ carbon atom with a +1 charge, +1 hydrophobicity, +1 hydrogen bond donor, and +1 hydrogen bond acceptor at each lattice and grid spacing 2.0 Å calculate the CoMSIA descriptors and the Gaussian function calculate the similarity indices between the molecules and the probe atom.

For the generation of the CoMSIA 3D-QSAR model, partial least-squares (PLS) algorithm was subjected to linearly correlate the CoMSIA fields to the cytotoxicity of the Knoevenagel-type indoles. The predictive ability of the constructed PLS model was judged by the cross-validation leave-one-out (LOO) method. The external predictability of the model was done by calculating the R^2_{pred} on the test set compounds.²⁶

Molecular Docking

Molecular docking was subjected to understand the binding modalities of investigated Knoevenagel-type indole molecules toward its putative target. The docking experiments were performed with the AutoDock Vina software.²⁸ The crystallographic structures of tubulin receptor (pdb 1SA0)²⁹ were taken from the RCSB Protein Data Bank (<http://www.pdb.org>). The receptors were freed from water molecules and inbound ligand by Discovery Studio 3.5 Visualizer (DS Visualizer 3.5, Accelrys Software Inc., San Diego, California, USA). From the

AutoDock Tools (ADT) the polar hydrogen atoms were added to the protein models to relax the conformation from close contacts. The grid box was selected for protein centred at the crystal structure of the corresponding inhibitors. The grid dimensions were chosen as with a spacing of 1 Å. The preparation of the ligand structure prior to docking was performed by adding polar hydrogen atoms, Gasteiger partial charges and the torsions were allowed to rotate during docking.³⁰ The docked poses were visualized and the 3D images were constructed by using the PyMOL software (The PyMOL Molecular Graphics System, Version 1.0r1, Schrödinger, LLC).

Results and Discussion

2D-QSAR study

The indole-based derivatives³ were used to build the chemometric models. The dataset was divided on the basis of the *k*-means clustering analysis (*k*-MCA) and the model was generated on the training set molecules and validated against the test set molecules. The correlation matrix of response variable and descriptors used in 2D-QSAR model was demonstrated in the supplementary materials. The regression analysis was carried out and the best equation of having six variables obtained is:

$$pEC_{50} = 3.885 (\pm 0.404) + 0.564 (\pm 0.083) H\text{-}AcceptorCount - 0.991 (\pm 0.257) H\text{-}DonorCount + 0.261 (\pm 0.103) SaasCE\text{-}index + 0.309 (\pm 0.078) G_O_O_4 - 0.359 (\pm 0.049) T_O_2_5 + 0.069 (\pm 0.012) G_2_2_5. \quad \text{Eq. (3)}$$

$$n_{Train} = 30; R = 0.917; R^2 = 0.842; R^2_A = 0.801; F(6, 23) = 20.395; SEE = 0.334; Q^2 = 0.722; PRESS = 4.497; SDEP = 0.387; rm^2_{(LOO)} = 0.683; \Delta rm^2_{(LOO)} = 0.110; n_{Test} = 9; R^2_{pred} = 0.799;$$

$$rm^2_{(test)} = 0.772; \Delta rm^2_{(test)} = 0.043; rm^2_{(overall)} = 0.697; \Delta rm^2_{(overall)} = 0.058; cR^2_p = 0.757; r^2 = 0.852; |r^2_{\theta} - r^2_{\theta'}| = 0.024; (r^2 - r^2_{\theta}) / r^2 = 0.028; (r^2 - r^2_{\theta'}) / r^2 = 0.000.$$

From the equation (3), it may be predicted that the variance of the equation is 80.1%, whereas the LOO-predicted variance 72.2%. Eq. (3) has good R^2_{pred} values of 0.799 suggest robust and significant external validation. This eq. (3) successfully passed all the validation test including rm^2 matrices, randomization test ($cR^2_p = 0.757$) and also the criteria as suggested by Golbraikh and Tropsha acceptable model criteria.^{5,23,25} The Observed (Obs) and Predicted (Pred) activities of training set and test set compounds of eq. (3) was given in Supplementary files. The observed versus predicted activities as well as the contribution charts of the descriptors as per eq. (3) were graphically shown in Figure 2.

[Figure 2 may be placed here]

The QSAR eq. (3) demonstrated that increasing values of the parameters such as *H-AcceptorCount* (number of hydrogen bond acceptor atoms), *SaasCE-index* (Electrotopological state index for number of carbon atoms connected with one single bond along with two aromatic bonds), *G_O_O_4* (the count of number of oxygen atoms separated from another oxygen atoms by four bond distance) and *G_2_2_5* (the count of number of double bonded atoms separated from another double bonded atoms by five bond distance) values may enhance the activity and increasing the *H-DonorCount* (number of hydrogen bond donor atoms), *T_O_2_5* (the count of number of oxygen atoms separated from double bonded atoms by five bond distance) may diminish the activity of the investigated indole-based compounds. The positive relationship of hydrogen bond acceptor atoms count descriptor reveals the importance of hydrogen bonding acceptor atoms for potent molecule design. As like compound **39** ($EC_{50} = 0.2 \mu\text{M}$; *H-*

AcceptorCount value of 6) is more potent than compound **1** ($EC_{50} = 36 \mu\text{M}$) and compound **13** ($EC_{50} = 6 \mu\text{M}$) having *H-AcceptorCount* value of 2 and 3 respectively. The higher values of *H-AcceptorCount* are observed in the presence of sulphonyl substituent on Y position (Figure 1) of the Knoevenagel-type indoles. Compounds possessing oxygen atoms as ester, keto and sulphonyl functions (compounds **20-22**, **30-38**) may be capable of accepting hydrogen bonds from the receptor and therefore, these compounds are found to be highly active as far as the cytotoxicity is concerned. Second descriptor *H-DonorCount* is inversely proportional to the biological activity data, thereby, cytotoxicity is augmented by minimizing hydrogen bonding donor atoms in these Knoevenagel-type indoles [compound **28** ($EC_{50} = 1.5 \mu\text{M}$) > compound **13** ($EC_{50} = 6 \mu\text{M}$) has *H-DonorCount* value of 1 and 2 respectively]. Compounds **1-3** are less potent analogs having lower values of *SaasCE-index*. Thereby increasing the *SaasCE-index* value may promote the cytotoxicity of the molecules having aromatic group at R position (Figure 1). The values of the alignment independent descriptor *G_O_O_4* decreases when more number of oxygen atoms is present at four bond distance. Compounds **30-33** having higher *G_O_O_4* values and they shows better cytotoxicity. Another alignment independent descriptor *T_O_2_5* negatively contributes in the eq. (3). The lower *T_O_2_5* values are observed generally in presence of the cyano group on the X position of the parent ring (Figure 1). Less number of oxygen atoms separated from double bonded atoms by five bond distances is found for compounds whose X portion is occupied by cyano group. Thus, the presence of the cyano function gives the lowest value of *T_O_2_5*, and therefore, increasing the cytotoxicity potency (Compound **6**, **10** vs **5,26**). The descriptor *G_2_2_5* suggests that substitution from the methyl to phenyl group will augment the biological activity of these indole-based compounds. This result may be easily understood by as evidenced in compound **1** ($EC_{50} = 36 \mu\text{M}$) bearing methyl group (*G_2_2_5* value 3.496) was less active

than compound **36** ($EC_{50} = 1 \mu\text{M}$) containing phenyl group ($G_{2_2_5}$ value 19.924) have almost structural similarity except the R substitution (Figure 1). Similarly, this phenomena was also evidenced when comparing compound **5** ($G_{2_2_5}$ value 27.298) and **31** ($G_{2_2_5}$ value 45.494) where the replacement of methyl group with phenyl group would result in higher biological activities ($EC_{50} \text{ 31 } \mu\text{M} < 1 \mu\text{M}$).

3D-QSAR study

kNN-MFA models

The similar training set (30 compounds), test set (9 compounds) was used 3DQSAR methods (Table 1). Molecular alignment of all compounds was performed on compound 35 as the reference best active compound. The aligned geometry of all these molecules was shown in Figure 3.

[Figure 3 may be placed here]

3D QSAR kNN-MFA models were generated by stepwise forward (SWF) selection and simulated annealing (SA) methods. Statistical parameters of the 3D-QSAR models were shown in Table 2.

[Table 2 may be placed here]

The kNN-MFA models derived by the SWF and the SA methods were shown in eq. (4) and eq. (5) respectively.

$$pEC_{50} = S_{1359} (4.5539, 4.8213) - S_{888} (-0.1365, -0.1351) - S_{898} (-0.0655, -0.0646) \quad \text{Eq. (4)}$$

$n_{Train} = 30$; $n_{Test} = 9$; k Nearest Neighbour = 2; $df = 26$; $Q^2 = 0.842$; $R_{pred}^2 = 0.887$; $Q^2_{SE} = 0.297$ and $R_{pred-SE}^2 = 0.289$.

$$pEC_{50} = S_952 (30.0000, 30.0000) - E_1289 (-0.0173, -0.0089) \quad \text{Eq. (5)}$$

$n_{Train} = 30$; $n_{Test} = 9$; k Nearest Neighbour = 2; $df = 27$; $Q^2 = 0.771$; $R_{pred}^2 = 0.733$; $Q^2_{SE} = 0.358$ and $R_{pred-SE}^2 = 0.445$.

The stepwise forward kNN-MFA model, i.e., eq. (4) yielded LOO cross-validation (Q^2) and predictive squared correlation coefficient (R_{pred}^2) of 0.842 and 0.887 respectively. This eq. (4) shows the contribution of the descriptors such as S_1359, S_888 and S_898 (the steric field energy of interactions between the methyl probe atom and compounds at their corresponding spatial grid points of 1359, 888, 898 respectively) for these indole derivatives. The contribution plot of model 4 for steric and electrostatic interactions was shown in Figure 4.

[Figure 4 may be placed here]

In eq. (4), it is observed that the positive steric interaction field at lattice point 1359 may be conducive for the biological activity. It is observed that the steric field is situated near to the *p*-phenyl substitutions attached to the indole moiety (Figure 1). Therefore, it may be assumed that bulky substitutions at this region may enhance the biological potential of these indoles by inducing favourable steric interaction with the receptor. Compounds with phenyl substitution (compounds **6-33**) at R position (Figure 1) are better active than the unsubstituted (compounds **1-3**) or methyl substituted analogs (compounds **4-5**) at this position. Moreover, compounds **34**, **35**, **38** and **39** having *p*-methoxy substitution at the phenyl ring (position R in Figure 1) are far better active than other compounds in this series. Whereas, contribution of steric interaction at lattice

points 888 and 898 respectively may contribute negatively to the biological activity according to eq. (4). Therefore, bulky substitutions at these positions may not be favourable for the activity. Compounds having bulky substitutions such as $\text{PO}(\text{OEt})_2$ and $\text{PO}(\text{OiPr})_2$ group (compounds **2** and **16**), COOEt group (compounds **3**, **25** and **27**), CONH_2 group (compound **13**) and COPip group (compound **14**) at X and Y position (Figure 1) may involve some unfavourable steric hindrance with the receptor and these compounds were found to possess lesser activity than other compounds in this series. Therefore, less bulky substitutions are preferable at X and Y position (Figure 1).

In simulated annealing kNN-MFA model, i. e., eq. (5) resulted in LOO cross-validation (Q^2) and predictive squared correlation coefficient (R_{pred}^2) of 0.771 and 0.733 respectively. This eq. (5) shows the contribution of the descriptors such as S_952 and E_1289 (the steric and electrostatic field energy of interactions between the methyl probe atom and compounds at their corresponding spatial grid points of 952 and 1289 respectively) for these indole derivatives. The contribution plot as per eq. (5) for steric and electrostatic interactions was shown in Figure 4. Here, it is observed that the positive steric interaction field at lattice point 952 and negative electrostatic interaction at lattice point 1289 may be conducive for the biological activity. Moreover, it is observed that the positive steric field interactions were situated near the substituted phenyl ring at position R (Figure 1). This supports the results obtained from stepwise forward kNN-MFA model (eq. 4). The negative electrostatic interaction suggested that decreasing the charge or the electrostatic potential at lattice point 1289 may contribute positively to the biological activity. It is observed that the region was located near the X and Y substitutions. Therefore, it may be assumed from eq. 4 and 5 that less electrostatic as well as less steric features at this position may be conducive to the biological activity. The Observed (Obs)

and Predicted (Pred) activities of training set and test set compounds of 3D-QSAR kNN-MFA models were given in Supplementary files and graphically shown in Figure 5.

[Figure 5 may be placed here]

3D-QSAR CoMSIA model

For developing 3D-QSAR CoMSIA models, again, all these molecules were aligned on the best active reference compound **35**. The CoMSIA alignment was shown in Figure 3. The CoMSIA model showed a LOO-cross validation Q^2 of 0.524 with 4 components and R^2 value of 0.960. The validation parameters as per the CoMSIA model were shown in Table 2. The CoMSIA R^2_{pred} value was 0.872 that was statistically reliable and significant. Therefore, these parameters indicated validated internal predictability of these indole analogs internally and externally. The CoMSIA contour maps for the best active (compound **35**) and the lowest active (compound **3**) compound was shown in Figure 6.

[Figure 6 may be placed here]

The favourable steric fields (shown as green contour) exist near R substitution of the indole-based derivatives suggesting that the introduction of the phenyl or a *p*-anisyl group to replace the methyl or hydrogen group could be more beneficial for biological activity (compounds **6-33** vs compounds **1-5**). These results are in agreement with the results obtained from kNN-MFA models where S_952 (from SA-kNN-MFA), S_1359 (from SWF-kNN-MFA) bear positive value and indicated the importance of bulky group in proximity of the R position (Figure 1). Previously Halder *et. al.* reported the QSAR studies on of some 2-phenylindole 3-carbaldehydes derivatives and showed also the importance of phenyl group for the anticancer activity.³¹⁻³² Regarding the

steric unfavourable fields (shown as yellow contour) that were positioned around X and Y substitution (Figure 1) of the parent residue, recommended less bulky group is tolerable to maintain cytotoxic potency of these compounds. In a recent study, Adhikari *et. al.* reported that bulky substituents at the 3 positions of phenylindole are detrimental for the anticancer activity.⁵ Similarly in our SWF-kNN-MFA model, the static field descriptor S_888, S_898 suggest that less bulky group is beneficial in proximity of the Y group (Figure 1) to increase the anti-proliferative activity of these indole analogs. This result is in close agreement with the experimental data where replacement of cyanide or acetyl group with acetophenone moiety at X position (Figure 1) may lead to poor activity (compounds **6-12**, **20-22** vs. compounds **3**, **23-28**). Moreover, the attachment of PO(OEt)₂, PO(OiPr)₂ group at Y position (Figure 1) is found to be detrimental for the anti-cancer activity (compounds **2** and **16**).

The electrostatic blue contour map especially enlightened the effect of the electron with-drawing compound in proximity of the Y substitution (Figure 1). The reliability of this result was also supported by the SA-kNN-MFA electrostatic descriptor E_1289. It means cyano, ester, keto and sulphonyl substituent at Y position (Figure 1) promotes the cytotoxic potential (compounds **6-12**, **34-39** etc). This is also correlated with a previous comparative molecular field analysis (CoMFA) study of 2-phenylindole derivatives by Liao and co-workers.³³ On the other hand, the presence of partially negative charged functions is suitable around the R position (Figure 1) as suggested by the macro red contours of the CoMSIA contour maps (Figure 6). This result seemed to signify the importance of the *para*-methoxy substitution on the phenyl ring of the R position and matching with the previous study.³³

The unfavourable hydrophobic field (cyan contour) cover the R' substitution and cover a part of the phenyl ring of Y substitution (Figure 1) of the indole-based derivatives. Similarly, the

hydrogen bond acceptor field was found to be unfavourable near the R' substitution (Figure 1). The favourable hydrogen bond acceptor field was found near the sulphonyl function (Figure 6) which in turn, supports the 2D-QSAR result. No hydrogen bond donor field was found during CoMSIA analysis and validation. Therefore, it may be assumed that hydrogen bond acceptor field may play important role than hydrogen bond donor field for the cytotoxic potential of these compounds. Our study The Observed (Obs) and Predicted (Pred) activities of training set and test set compounds of 3D-QSAR CoMSIA model was given in Supplementary files and graphically shown in Figure 5.

Molecular docking study

Tubulin has been preliminary identified as the biological target for this kind of molecules.³ In order to identify how these molecules can bind to tubulin we performed a docking study with the tubulin receptor (pdb 1SA0).²⁹ To validate the docking protocol we performed re-docking with the crystallographic ligand colchicine into the active site of tubulin and have observed that re-docked colchicine and crystallographic colchicine are almost at similar position in the active site of the tubulin (Figure 7). After that we have docked the most potent ligand i.e. compound no **35** and a complete overview of the receptor-inhibitor binding interaction is presented in the Figure 7. The figure suggests that the compound can nicely bind in the colchicine binding site and there are several interactions between the compound and the active site of the enzyme. The residues Leu 248, Leu 255, Lys 254, Val 238, Val 318, Met 259, Val 315 etc. are important for the hydrophobic interaction with the phenyls of the ligand. The cation- π interaction is formed between side chain of Lys 352 and the molecular frame of 1H indole similar to the earlier report.³³ A hydrophobic pocket comprising Leu 248, Ala 250 and Leu 255 found near R substitution pointing out that the phenyl or *p*-anisyl instead of methyl/hydrogen are favourable

for activity which is correlated with our COMSIA study. The R' position is very close to the amino acids e.g. Asn 350, Lys 352, Thr 353 residues making it unfavourable for hydrophobic substituent for the anticancer activity. The substituents X and Y are very close to the β -sheets, Tyr312 - Arg320 and Val351 - Cys356 and there is no enough space to accommodate the bulky groups (compounds **6-12**, **20-22** vs. compounds **3**, **23-28**).

[Figure 7 may be placed here]

Conclusion

In this work, multiple molecular modelling tools were applied to explore the structural and physicochemical requirements of some unconventional Knoevenagel-type indole derivatives for their cytotoxic potential. The 2D and 3D-QSAR models are statistically significant and well validated. The 2D-QSAR model suggested that increasing number of hydrogen bond acceptor and decreasing number of hydrogen bond donor feature may be conducive to the activity which in turn supported by the CoMSIA study and correlate the SAR data as well. CoMSIA study suggested the importance of favourable steric feature around the aromatic substitutions attached to the indole ring which in turn validated the kNN-MFA 3D-QSAR results. This ligand-based modelling approach may help to understand the structural and physicochemical information that could be beneficial in designing novel drug like molecules of this class in anti-cancer drug designing. Structural requirements of these indole analogs for being capable of higher cytotoxic potential are shown in Figure 8.

[Figure 8 may be placed here]

Abbreviations

n_{Train} : number of train set compounds,

n_{Test} : number of test set compounds,

R^2_A : adjusted squared correlation coefficient (R^2),

SEE : standard error of estimation,

$PRESS$ = predicted residual sum of square,

$SDEP$ = and standard deviation of error of prediction,

df : degree of freedom,

r^2 : the squared correlation coefficient between the observed vs predicted response values of the test set compounds,

r^2_o : the corresponding values for regression through origin (observed vs. Predicted),

$r^2_{o'}$: the corresponding values for regression through origin (predicted vs. observed),

Q^2_SE : standard error of the LOO cross-validation parameter (Q^2),

$R^2_{pred_SE}$: standard error of the predictive squared correlation coefficient (R^2_{pred}).

Acknowledgements

S.G is thankful for the Start up grant from University Grants Commission (UGC), New Delhi (No.F.30-106/2015-BSR). One of the authors (S.A.A) is grateful to All India Council for Technical Education (AICTE), New Delhi for awarding post Graduate fellowship. The author (N.A) thanks University Grants Commission (UGC), New Delhi for providing Rajiv Gandhi National Fellowship (RGNF). We are also thankful to the authority of Department of Pharmaceutical Sciences, Dr. Harisingh Gour University, Sagar, India and Department of Pharmaceutical Technology, Jadavpur University, Kolkata, India for providing research facility and encouragement.

Supportive/Supplementary Material:

1. Supplementary table T1. Correlation matrix of response variable and descriptors used to develop 2D-QSAR model.
2. Supplementary table T2. Observed (Obs) and Predicted (Pred) activities of compounds for 2D-QSAR eq. (3).
3. Supplementary table T3. Observed (Obs) and Predicted (Pred) activities of training set compounds of 3D-QSAR models.
4. Supplementary table T4. Observed (Obs) and Predicted (Pred) activities of test set compounds of 3D-QSAR models.

Draft

References

- (1) De Sa Alves, F. R.; Barreiro, E. J.; Fraga, C. A. *Mini Rev. Med. Chem.*, **2009**, *9*, 782.
- (2) Gribble, G. W. *J. Chem. Soc.* **2000**, *1*, 1045.
- (3) Spallarossa, A.; Caneva, C.; Caviglia, M.; et al. *Eur. J. Med. Chem.* **2015**, *102*, 648.
- (4) Mondal, C.; Halder, A. K.; Adhikari, N.; Saha, A.; Saha, K. D.; Gayen, S.; Jha, T. *Eur. J. Med. Chem.* **2015**, *90*, 860.
- (5) Adhikari, N.; Halder, A. K.; Saha, A.; Saha, K. D.; Jha, T. *Tox. In Vit.* **2015**, *29*, 1392.
- (6) Adhikari, N.; Halder, A. K.; Mondal, C.; Jha, T. *Med. Chem. Res.* **2014**, *23*, 3096.
- (7) Mondal, C.; Halder, A. K.; Adhikari, N.; Jha, T. *Mol. Divers.* **2014**, *18*, 655.
- (8) Adhikari, N.; Halder, A. K.; Mondal, C.; Jha, T. *Med. Chem. Res.* **2013**, *22*, 6029.
- (9) Debnath, B.; Gayen, S.; Bhattacharya, S.; Samanta, S.; Jha, T. *Bioorg. Med. Chem.* **2003**, *11*, 5493.
- (10) Debnath, B.; Gayen, S.; Samanta, S.; Basu, A.; Ghosh, B.; Jha, T. *Ind. J. Chem.* **2006**, *45*, 93.
- (11) Gayen, S.; Debnath, B.; Samanta, S.; Jha, T. *Bioorg Med Chem.* **2004**, *12*, 1493.
- (12) Jana, D.; Halder, A. K.; Adhikari, N.; Maiti, M. K.; Mondal, C.; Jha T. *Med. Chem. Commun.* **2011**, *2*, 840.
- (13) Adhikari, N.; Maiti, M. K.; Jha, T. *Bioorg. Med. Chem. Lett.* **2010**, *20*, 4021.
- (14) Adhikari, N.; Maiti, M. K.; Jha, T. *Eur. J. Med. Chem.*, **2010**, *45*, 1119.
- (15) Halder, A. K.; Adhikari, N.; Maiti, M. K.; Jha T. *Eur. J. Med. Chem.*, **2010**, *45*, 1760.
- (16) Alam, S. M.; Samanta, S.; Halder, A. K.; Basu, S.; Jha, T. *Can. J. Chem.* **2007**, *85*, 1053.
- (17) Basu, A.; Gayen, S.; Samanta, S.; Panda, P.; Srikanth, K.; Jha, T. *Can. J. Chem.* **2006**, *84*, 458.

- (18) VLife Sciences Technologies Pvt. Ltd. *VLife MDS 4.4*; VLife Sciences Technologies Pvt. Ltd.: Pune, **2015**; available from <http://www.vlifesciences.com>.
- (19) *SYBYL-X 2.0*. Tripos Inc., St. Louis. **2015**; available from <http://www.tripos.com>.
- (20) *Chem 3D Pro Version 5.0* and *Chem Draw Ultra Version 5.0* are software programs developed by Cambridge Soft Corporation, U. S. A.
- (21) Tropsha, A. In: *Burger's Medicinal Chemistry and Drug Discovery*, D. J. Abraham, (Eds.); John Wiley and sons Inc., New Jersey, **2003**; Vol. 1, pp. 49-75.
- (22) Tetko, I. V.; Tanchuk, V. Y.; Villa, A. E. *J. Chem. Inf. Comput. Sci.* **2001**, 41, 1407.
- (23) Golbraikh, A.; Tropsha, A. *J. Mol. Graph. Model.* **2002**, 20, 269.
- (24) Ojha, P. K.; Mitra, I.; Das, R. N.; Roy, K. *Chemo. Intell. Lab. Syst.* **2011**, 107, 194–205.
- (25) Tropsha, A.; Gramatica, P.; Gombar, V. K. *QSAR Comb Sci.* **2003**, 22, 69.
- (26) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. *J. Am. Chem. Soc.* **1988**, 110, 5959.
- (27) Noolvi, M. N.; Patel, H. M. *J. Saudi Chem. Society.* **2013**, 17, 361.
- (28) Trott, O.; Olson, A. J. *J. Comp. Chem.* **2010**, 31, 455.
- (29) Ravelli, R. B.; Gigant, B.; Curmi, P. A.; Jourdain, I.; Lachkar, S.; Sobel, A.; Knossow, M. *Nature* **2004**, 428, 198.
- (30) Lv, P. C.; Li, H. Q.; Sun, J.; Zhou, Y.; Zhu, H. L. *Bioorg. Med. Chem.* **2010**, 18, 4606.
- (31) Halder, A. K.; Adhikari, N.; Jha T. *Bioorg. Med. Chem. Lett.* **2009**, 19, 1737.
- (32) Halder, A. K.; Adhikari, N.; Jha T. *Chem. Biol. Drug. Des.* **2010**, 75, 204.
- (33) Liao, S. Y.; Qian, L.; Miao, T. F.; Lu, H. L.; Zheng, K. C. *Eur. J. Med. Chem.* **2009**, 44, 2822.

Figure 1. General structure of Knoevenagel-type indole derivatives with arbitrary numbering.

Figure 2. (A) Graphical representation of observed (Obs) versus predicted (Pred) activities of Knoevenagel-type indole derivatives for 2D-QSAR model; (B) Contribution charts of the descriptors used to develop the 2D-QSAR model.

Figure 3. Molecular alignment of unconventional Knoevenagel-type indoles derivatives for developing kNN-MFA model (A) and (B) CoMSIA model.

Figure 4. Contribution plot for steric and electrostatic interactions in SWF- kNN MFA model (A) and SA-kNN MFA model (B).

Figure 5. Graphical representation of Observed (Obs) versus Predicted (Pred) activities of Knoevenagel-type indole derivatives for 3D-QSAR models [(A) SWF- kNN MFA model, (B) SA-kNN MFA model, (C) CoMSIA model]

Figure 6. CoMSIA contour map in transparent mesh for the best active compound **35** (A) and the lowest active compound **3** (B). [Green: Steric favorable; Yellow: Steric unfavorable; Blue: Electropositive favorable; Red: Electropositive unfavorable; Magenta: Hydrophobic favorable; Cyan: Hydrophobic unfavorable; Orange: Hydrogen-bond acceptor favorable; White: Hydrogen bond acceptor unfavorable].

Figure 7. Binding conformation of the crystal colchicine (red) and the redocked colchicine (green) at the tubulin binding site (A). Docking conformation for the best active compound **35** at the colchicine binding site (B)

Figure 8. Structural requirements of Knoevenagel-type indole derivatives for anticancer activity.

Table 1. Structures of the Knoevenagel-type indoles together with the cytotoxicity data.

Cpd ^a	R	R'	X	Y	EC ₅₀ (μM)	pEC ₅₀
1	H	H	CN	CN	36.00	4.44
2 ^b	H	H	CN	PO(OEt) ₂	80.00	4.10
3	H	H	<i>p</i> -NO ₂ PhCO	COOEt	100.00	4.00
4	Me	H	CN	CN	56.00	4.25
5	Me	H	COOMe	SO ₂ Ph	31.00	4.51
6	Ph	H	CN	CN	0.40	6.40
7	Ph	H	CN	COC(Me) ₃	0.50	6.30
8 ^b	Ph	H	CN	COPh	0.50	6.30
9	Ph	H	CN	<i>p</i> -ClPhCO	0.40	6.40
10 ^b	Ph	H	CN	2-Theonyl	0.40	6.40
11	Ph	H	CN	COOEt	1.70	5.77
12	Ph	H	CN	COO(CH ₂) ₂ OMe	1.20	5.92
13	Ph	H	CN	CONH ₂	6.00	5.22
14 ^b	Ph	H	CN	COPip	9.60	5.02
15	Ph	H	CN	<i>p</i> -ClPhNHCO	1.70	5.77
16	Ph	H	CN	PO(OiPr) ₂	11.00	4.96
17 ^b	Ph	H	CN	SO ₂ Me	0.30	6.52
18	Ph	H	CN	SO ₂ Ph	0.40	6.40
19 ^b	Ph	H	CN	<i>p</i> -MePhSO ₂	0.40	6.40
20	Ph	H	COMe	COMe	1.50	5.82
21	Ph	H	COMe	SO ₂ Me	0.50	6.30
22	Ph	H	COMe	SO ₂ Ph	0.40	6.40
23 ^b	Ph	H	COPh	COMe	2.00	5.70
24	Ph	H	COPh	COPh	8.90	5.05
25	Ph	H	COPh	COOEt	7.00	5.15
26	Ph	H	COPh	SO ₂ Ph	3.00	5.52
27 ^b	Ph	H	<i>p</i> -ClPhCO	COOEt	9.00	5.05
28	Ph	H	<i>p</i> -NO ₂ PhCO	COOEt	1.50	5.82
29	Ph	H	2-Furoyl	COCF ₃	0.80	6.10
30 ^b	Ph	H	COOMe	SO ₂ Me	0.40	6.40
31	Ph	H	COOMe	SO ₂ Ph	1.00	6.00
32	Ph	H	COOEt	COMe	0.60	6.22
33	Ph	H	COOEt	COOEt	2.30	5.64
34	<i>p</i> -OMePh	H	CN	CN	0.30	6.52
35	<i>p</i> -OMePh	H	CN	SO ₂ Ph	0.20	6.70
36	Ph	OMe	CN	CN	1.00	6.00
37	Ph	OMe	CN	SO ₂ Ph	1.30	5.89
38	<i>p</i> -OMePh	OMe	CN	CN	0.30	6.52
39	<i>p</i> -OMePh	OMe	CN	SO ₂ Ph	0.20	6.70

^aCompound number; ^bCompounds considered as the test set molecules

Table 2. Statistical parameters of the 3D-QSAR models.

Entry	Parameter	SWF-kNN-MFA	SA-kNN-MFA	CoMSIA
1	n_{Train}	30	30	30
2	n_{Test}	9	9	9
3	k Nearest Neighbour	2	2	-
4		26	27	-
5		-	-	0.960
6	SEE	-	-	0.161
7	Q^2	0.842	0.771	0.524
8	Q^2_{SE}	0.297	0.358	-
9	R^2_{pred}	0.887	0.733	0.872
10	$R^2_{pred-SE}$	0.289	0.445	-
11	F value	-	-	149.170

Draft

(A)

(B)

SWF-kNN-MFA model

SA-kNN-MFA model

CoMSIA model

<https://mc.manuscriptcentral.com/cjc-pubs>

✓ hydrophobic substituent unfavourable,
OCH₃ group is suitable for cytotoxicity
(CoMSIA and 2D-QSAR model)

✓ cyano, ester,
keto, sulphonyl
substituent at Y
position promotes
the cytotoxic
potential (kNN-
MFA, CoMSIA
and 2D-QSAR
model)

✓ attachment of
Electron with-
drawing cyanide
group favour
biological potency
(kNN-MFA,
CoMSIA and 2D-
QSAR model)

✓ bulky phenyl/ *p*-anisyl group
substituent is favourable for
anti-cancer agent (kNN-MFA,
CoMSIA and 2D-QSAR model)