Scientific paper

Prediction of Anticancer Activity of 2-phenylindoles: Comparative Molecular Field Analysis Versus Ridge Regression using Mathematical Molecular Descriptors

Subhash C. Basak, Qianhong Zhu and Denise Mills

University of Minnesota Duluth, Natural Resources Research Institute, 5013 Miller Trunk Highway, Duluth, MN 55811, USA

** Corresponding author: E-mail: sbasak@nrri.umn.edu*

Received: 16-04-2009

This paper is dedicated to Professor Milan Randić on the occasion of his 80th birthday

Abstract

Topological indices (TIs) and atom pairs (APs) were used to develop quantitative structure-activity relationships (QSARs) for anticancer activity for a set of 43 derivatives of 2-phenylindole. Results show that QSARs formulated using TI+AP outperform those using either TI or AP alone. The q^2 of the ridge regression model using TI+AP was 0.867 as compared to 0.705 reported in the literature using the comparative molecular field analysis (CoMFA) method.

Keywords: Anticancer activity, Phenylindole, Tubulin, Colchicine site inhibitors (CSIs), Comparative molecular field analysis (CoMFA), Mathematical molecular descriptors

1. Introduction

Tubulins consist of a small group of globular proteins with approximate molecular weight of 55 kilodaltons. The most common members of the tubulin family are α-tubulin and β-tubulin. Microtubules are assembled as dimers of α- and β-tubulin subunits.[1] Microtubule is the generic name of a class of subcellular components that occur in a wide variety of eukaryotic cells. Such structures are straight cylinders, 240 ± 20 Å in diameter, with a hollow 150 Å core. They have diverse biochemical functions which include chromosome movements in cell division, intracellular transport of materials, development and maintenance of cell form, cellular motility, and sensory transduction. It is well known that the disruption of microtubules by antimitotic drugs or physical factors results in disruption of cellular function.[2]

Various tubulin binding ligands with antimitotic and anticancer properties have been reported in the literature.[3–6] Regarding the binding sites of the various ligands, these can be classified into three main groups: those that bind tubulin at the colchicine-binding site; those that bind at the vinblastine site, and those that bind at the taxol site.

The inhibition of microtubule formation via tubulin polymerization results in mitotic arrest which, in turn, promotes vascular disruption, leading to cell death by apoptosis. Hence, tubulin has emerged as a popular target for anticancer drug design.[7, 8]

Von Angerer et al. synthesized a group of 2-phenylindole derivatives and determined their anticancer activities in human breast cancer cells.[9–11] One of their critical observations was that these compounds prevent the polymerization of the α/β -tubulin dimers to functional microtubules by binding to the colchicine-binding site and all have pronounced cytotoxicity, indicating their good potential as a new class of anticancer drugs. Consequently, there has been a lot of interest in understanding the structural basis of the anticancer activity of 2-phenylindoles using quantitative structure-activity relationship (QSAR) modeling. In fact, Liao et. al.[12] applied the comparative molecular field analysis (CoMFA) approach to a set of 43 analogs of 2-phenylindole with reasonable results. In our previous studies we found that mathematical molecular descriptors, invariants of simple and weighted molecular graphs in particular, which can be calculated directly from chemical structure without the input of any other experimental data, can predict property/ bioactivity/toxicity of various congeneric and structurally diverse classes of chemicals.[13–24] So in this paper we carried out QSAR modeling on the set of 43 2-phenylindoles using a diverse collection of mathematical structural invariants.

2. Materials and Methods

2. 1. The Database

The 43 compounds used for the QSAR models in this study were taken from the published work of von Angerer and his coworkers.[9–11] Liao *et al.*[12] carried out a CoMFA type of QSAR using this set of compounds. The anticancer activity of the 43 2-phenylindole derivatives was measured as the level of cytotoxicity against human breast cancer cell line MDA-MB 231. The range of IC_{50} values was 5.5 to 720 nM, more than two orders of magnitude between the most and least potent derivatives. We used pIC_{50} values of the compounds ($pIC_{50} = -\log IC_{50}$) as dependent variable in our models. The structural formula of the studied compounds is shown in Fig 1. The structure of each compound and its bioactivity are listed in Table 1.

Table 1. Structures and anticancer activities against human breast cancer cell line MDA-MB 231

No.	R_1	R_2	R_3	X	IC_{50}(nm)	pIC_{50}
1	H	H	H	$C(CN)_2$	430	6.367
2	H	H	OCH_3	$C(CN)_2$	720	6.143
3	H	OCH_3	OCH_3	$C(CN)_2$	590	6.229
4	OCH_3	H	OCH_3	$C(CN)_2$	260	6.585
5	H	F	OCH_3	$C(CN)_2$	400	6.398
6	F	H	OCH_3	$C(CN)_2$	280	6.553
7	OCH_3	H	CH_3	$C(CN)_2$	180	6.745
8	H	CH_3	OCH_3	$C(CN)_2$	280	6.553
9	Cl	CH_3	OCH_3	$C(CN)_2$	75	7.125
10	H	n-Pr	OCH_3	$C(CN)_2$	83	7.081
11	H	i-Pr	OCH_3	$C(CN)_2$	210	6.678
12	H	n-Bu	OCH_3	$C(CN)_2$	26	7.585
13	H	n-Pentyl	OCH_3	$C(CN)_2$	42	7.377
14	H	n-Hexyl	OCH_3	$C(CN)_2$	46	7.337
15	H	n-Bu	CH_3	$C(CN)_2$	65	7.187
16	H	n-Bu	CH_2CH_3	$C(CN)_2$	76	7.119
17	H	n-Bu	CF_3	$C(CN)_2$	56	7.252
18	H	n-Pentyl	CF_3	$C(CN)_2$	78	7.108
19	H	n-Hexyl	CF_3	$C(CN)_2$	150	6.824
20	H	OCH_3	OCH_3	O	260	6.585
21	OCH_3	H	OCH_3	O	35	7.456
22	F	H	OCH_3	O	59	7.229
23	H	F	OCH_3	O	540	6.268
24	Cl	H	OCH_3	O	27	7.569
25	Cl	CH_3	OCH_3	O	26	7.585
26	H	CH_3	OCH_3	O	86	7.066
27	H	Pr	OCH_3	O	20	7.699
28	H	n-Bu	OCH_3	O	6.7	8.174
29	H	sec-Bu	OCH_3	O	72	7.143
30	H	t-Bu	OCH_3	O	280	6.553
31	H	n-Pentyl	OCH_3	O	5.5	8.260
32	H	n-Hexyl	OCH_3	O	7.4	8.131
33	OCH_3	OCH_3	OCH_3	O	220	6.658
34	OCH_3	H	CH_3	O	31	7.509
35	H	CH3	CH_3	O	48	7.319
36	H	n-Bu	CH_3	O	34	7.469
37	H	n-Bu	CH_2CH_3	O	27	7.569
38	H	CH_2CH_3	n-Bu	O	300	6.523
39	H	n-Bu	CF_3	O	33	7.481
40	H	n-Pentyl	CF_3	O	42	7.377
41	H	n-Hexyl	CF_3	O	43	7.367
42	OCH_3	H	H	O	240	6.620
43	H	H	H	O	420	6.377

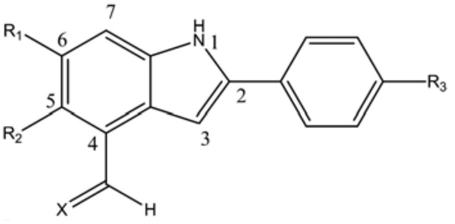

Fig.1. Molecular structure of 2-phenylindole derivatives

2. 2 Calculation of Molecular Descriptors

Two general classes of molecular descriptors were used as independent variables in the current study, namely, atom pairs (APs) and topological indices (TIs). The former are molecular substructures, while the latter are derived from graph theoretical methods. It is important to note that both types of descriptors are based solely on chemical structure.

An atom pair represents any two atoms in the molecule and includes information about their path-wise interatomic separation as well as the electronic character of the atoms. The method of Carhart *et al.* [25] was used in their calculation and defines an atom pair as a substructure consisting of two non-hydrogen atoms *i* and *j* and their interatomic separation:

<atom descriptor i> – <separation> – <atom descriptor j>

where <atom descriptor> contains information regarding atom type, number of non-hydrogen neighbors and the number of electrons. The interatomic separation is defined as the number of atoms traversed in the shortest bond-by-bond path containing both atoms. An example demonstrating the calculation of APs can be found in an earlier publication.[26] *APProbe* [27] was used to calculate the atom pairs for each molecule in the data set. In total, 354 APs were calculated for the data set.

In addition to the atom pairs, a set of 369 topological indices (TIs) was calculated using programs including *POLLY v2.3*,[28] *Triplet*[29] and *Molconn-Z v.3.5.*[30] They include path length descriptors, [31] path or cluster connectivity indices, [31, 32] neighborhood complexity indices,[33] valence path connectivity indices, [31] hydrogen bonding descriptors and electrotopological state indices.[34] Topological indices may be classified as either topostructural (TS) or topochemical (TC). The former encode information related to connectivity only, while the latter also encode chemical information such as atom and bond type. Table 2 provides a list of the topological indices calculated for this study, along with brief descriptions.

Prior to model development, any descriptor with a constant value for all, or nearly all, compounds within the data set was omitted. In addition, only one descriptor of any perfectly correlated pair (i.e., r = 1.0), as identified by the CORR procedure of the SAS statistical package[35] was retained. Subsequently, 248 TIs remained for use in the modeling study. Prior to modeling, the descriptors were

Table 2. Symbols, definitions and classification of topological indices

Topostructural (TS)	
I_D^W	Information index for the magnitudes of distances between all possible pairs of vertices of a graph
$\hat{I}_D^W$	Mean information index for the magnitude of distance
W	Wiener index = half-sum of the off-diagonal elements of the distance matrix of a graph
I^D	Degree complexity
H^V	Graph vertex complexity
H^D	Graph distance complexity
IC	Information content of the distance matrix partitioned by frequency of occurrences of distance *h*
M_1	A Zagreb group parameter = sum of square of degree over all vertices
M_2	A Zagreb group parameter = sum of cross–product of degrees over all neighboring (connected) vertices
${}^h\chi$	*Path connectivity index of order h = 0–10*
${}^h\chi_C$	*Cluster connectivity index of order h = 3–6*
${}^h\chi_{PC}$	*Path-cluster connectivity index of order h = 4–6*
${}^h\chi_{Ch}$	*Chain connectivity index of order h = 3–10*
P_h	*Number of paths of length h = 0–10*
J	*Balaban's J index based on topological distance*
nrings	*Number of rings in a graph*
ncirc	*Number of circuits in a graph*
DN^2S_y	Triplet index from distance matrix, square of graph order, and distance sum; operation y = 1–5
DN^21_y	Triplet index from distance matrix, square of graph order, and number 1; operation y = 1–5
$AS1_y$	Triplet index from adjacency matrix, distance sum, and number 1; operation y = 1–5
$DS1_y$	Triplet index from distance matrix, distance sum, and number 1; operation y = 1–5
ASN_y	Triplet index from adjacency matrix, distance sum, and graph order; operation y = 1–5

Topostructural (TS)	
DSN_y	Triplet index from distance matrix, distance sum, and graph order; operation y = 1–5
DN^2N_y	Triplet index from distance matrix, square of graph order, and graph order; operation y = 1–5
ANS_y	Triplet index from adjacency matrix, graph order, and distance sum; operation y = 1–5
$AN1_y$	Triplet index from adjacency matrix, graph order, and number 1; operation y = 1–5
ANN_y	Triplet index from adjacency matrix, graph order, and graph order again; operation y = 1–5
ASV_y	Triplet index from adjacency matrix, distance sum, and vertex degree; operation y = 1–5
DSV_y	Triplet index from distance matrix, distance sum, and vertex degree; operation y = 1–5
ANV_y	Triplet index from adjacency matrix, graph order, and vertex degree; operation y = 1–5
kp_0	Kappa zero
kp_1–kp_3	Kappa simple indices
Topochemical (TC)	
O	Order of neighborhood when IC_r reaches its maximum value for the hydrogen-filled graph
O_{orb}	Order of neighborhood when IC_r reaches its maximum value for the hydrogen-suppressed graph
I_{ORB}	Information content or complexity of the hydrogen-suppressed graph at its maximum neighborhood of vertices
IC_r	Mean information content or complexity of a graph based on the rth (r = 0–6) order neighborhood of vertices in a hydrogen-filled graph
SIC_r	Structural information content for rth (r = 0–6) order neighborhood of vertices in a hydrogen-filled graph
CIC_r	Complementary information content for rth (r = 0–6) order neighborhood of vertices in a hydrogen-filled graph
${}^{h}\chi^{b}$	Bond path connectivity index of order h = 0–6
${}^{h}\chi^{b}_{C}$	Bond cluster connectivity index of order h = 3–6
${}^{h}\chi^{b}_{Ch}$	Bond chain connectivity index of order h = 3– 6
${}^{h}\chi^{b}_{PC}$	Bond path-cluster connectivity index of order h = 4–6
${}^{h}\chi^{v}$	Valence path connectivity index of order h = 0–10
${}^{h}\chi^{v}_{C}$	Valence cluster connectivity index of order h = 3–6
${}^{h}\chi^{v}_{Ch}$	Valence chain connectivity index of order h = 3–10
${}^{h}\chi^{v}_{PC}$	Valence path-cluster connectivity index of order h = 4–6
J^B	Balaban's J index based on bond types
J^X	Balaban's J index based on relative electronegativities
J^Y	Balaban's J index based on relative covalent radii
AZV_y	Triplet index from adjacency matrix, atomic number, and vertex degree; operation y = 1–5
AZS_y	Triplet index from adjacency matrix, atomic number, and distance sum; operation y = 1–5
ASZ_y	Triplet index from adjacency matrix, distance sum, and atomic number; operation y = 1–5
AZN_y	Triplet index from adjacency matrix, atomic number, and graph order; operation y = 1–5
ANZ_y	Triplet index from adjacency matrix, graph order, and atomic number; operation y = 1–5
DSZ_y	Triplet index from distance matrix, distance sum, and atomic number; operation y = 1–5
DN^2Z_y	Triplet index from distance matrix, square of graph order, and atomic number; peration Y = 1–5
nvx	Number of non-hydrogen atoms in a molecule
nelem	Number of elements in a molecule
fw	Molecular weight
si	Shannon information index
totop	Total Topological Index *t*
sumI	Sum of the intrinsic state values *I*
sumdelI	Sum of delta-*I* values
tets2	Total topological state index based on electrotopological state indices
phia	Flexibility index (kp_1* kp_2/nvx)
Idcbar	Bonchev-Trinajstić information index
IdC	Bonchev-Trinajstić information index
Wp	Wienerp
Pf	Plattf
Wt	Total Wiener number
knotp	Difference of chi-cluster-3 and path/cluster-4
knotpv	Valence difference of chi-cluster-3 and path/cluster-4
nclass	Number of classes of topologically (symmetry) equivalent graph vertices

Topochemical (TC)	
NumHBd	Number of hydrogen bond donors
NumHBa	Number of hydrogen bond acceptors
SHCsats	E-State of C sp^3 bonded to other saturated C atoms
SHCsatu	E-State of C sp^3 bonded to unsaturated C atoms
SHvin	E-State of C atoms in the vinyl group, =CH–
SHtvin	E-State of C atoms in the terminal vinyl group, $=CH_2$
SHavin	E-State of C atoms in the vinyl group, =CH–, bonded to an aromatic C
SHarom	E-State of C sp^2 which are part of an aromatic system
SHHBd	Hydrogen bond donor index, sum of Hydrogen E–State values for –OH, =NH, $–NH_2$, –NH–, –SH, and #CH
SHwHBd	Weak hydrogen bond donor index, sum of C–H Hydrogen E-State values for hydrogen atoms on a C to which a F and/or Cl are also bonded
SHHBa	Hydrogen bond acceptor index, sum of the E-State values for –OH, =NH, $–NH_2$, –NH–, >N, –O–, –S–, along with –F and –Cl
Qv	General Polarity descriptor
$NHBint_y$	Count of potential internal hydrogen bonders (y = 2–10)
SHBinty	E–State descriptors of potential internal hydrogen bond strength (*y =2–10*)
ka_1–ka_3	Kappa alpha indices
Electrotopological State index values for atom types:	*SHsOH, SHdNH, SHsSH, SHsNH2, SHssNH, SHtCH, SHother, SHCHnX, Hmax Gmax, Hmin, Gmin, Hmaxpos, Hminneg, SsLi, SssBe, Sssss, Bem, SssBH ,SsssB, SssssBm, SsCH3, SdCH2, SssCH2, StCH, SdsCH, SaaCH, SsssCH, SddC, StsC, SdssC, SaasC, SaaaC, SssssC, SsNH3p, SsNH2, SssNH2p, SdNH, SssNH, SaaNH, StN, SsssNHp, SdsN, SaaN, SsssN, SddsN, SaasN, SssssNp, SsOH, SdO, SssO, SaaO, SsF, SsSiH3, SssSiH2, SsssSiH, SssssSi, SsPH2, SssPH, SsssP, SdsssP, SssssssP, SsSH, SdS, SssS, SaaS, SdssS, SddssS, SssssssS, SsCl, SsGeH3, SssGeH2, SsssGeH, SssssGe, SsAsH2, SssAsH, SsssAs, SdsssAs, SssssssAs, SsSeH, SdSe, SssSe, SaaSe, SdssSe, SddssSe, SsBr, SsSnH3, SssSnH2, SsssSnH, SssssSn, SsI, SsPbH3, SssPbH2,SsssPbH, SssssPb.*

standardized by autoscaling to zero mean and unit standard deviation.

2. 3. Statistical Analysis

Three regression methods that are appropriate when the number of descriptors exceeds the number of observations are ridge regression (RR),[36, 37] principal component regression (PCR),[38] and partial least squares (PLS) regression.[38, 39] These are shrinkage methods that avoid overfitting by imposing a penalty on large fluctuations of the estimated parameters. They are designed to utilize all available descriptors, as opposed to subset regression wherein variable selection is employed, and can be used with descriptors that are intercorrelated. RR, like PCR, transforms the descriptors to their principal components (PCs) and uses the PCs as descriptors. However, unlike PCR, RR retains all of the PCs, and žshrinks’ them differentially according to their eigenvalue.[36] As with PCR and RR, PLS also involves the creation of new axes in predictor space, however, they are based on both the independent and dependent variables.[40, 41] Statistical theory suggests that RR is the best of the three methods, and we have found in comparative studies that RR outperforms PCR and PLS in the vast majority of cases.[21, 39, 42–45] Therefore, we report only the ridge regression results in the current study. For the sake of brevity, we do not report the highly parameterized models, themselves, but rather the associated q^2 values, which are used to evaluate the predictive quality of the models. The q^2 is defined by:

$$q^2 = 1 - (PRESS \,/\, SS_{Total}) \qquad (1)$$

where *PRESS* is the prediction sum of squares and SS_{Total} is the total sum of squares. Unlike R^2, q^2 may be negative, indicative of a very poor model. Also, unlike R^2 which tends to increase upon the addition of any descriptor, q^2 will decrease upon the addition of irrelevant descriptors, providing a reliable measure of model quality.

The leave-one-out (LOO) method was used for model cross-validation. Unfortunately, it is a widely held belief that the use of a hold-out test set is always the best method of model validation. However, theoretic argument and empiric study[46] have shown that the LOO cross-validation approach is *preferred* to the use of a hold-out test set unless the data set to be modeled is very large. The drawbacks of holding out a test set include: 1) Structural features of the held out chemicals are not included in the modeling process, resulting in a loss of information, 2) Predictions are made on only a subset of the available compounds, whereas LOO predicts the activity value for all compounds, 3) There is no scientific tool that can guarantee similarity between the training and test sets, and 4)

Personal bias can easily be introduced in selection of the external test set. The reader is referred to Hawkins *et al.*[46] and Kraker *et al*.[47] for further discussion of proper model validation techniques.

The reader is cautioned to be critical of research studies which involve descriptor selection and cross-validation. In many such studies, the q^2 is obtained via a two-step process wherein a subset of descriptors is first selected, followed by cross-validation of the model which is developed based on those descriptors. This procedure results in an overly optimistic q^2 (termed "naïve q^2") which overestimates the predictive ability of the model.[47, 48] When using cross-validation and descriptor selection, it is essential that the descriptor selection step be included in the validation procedure. In doing so, the "true q^2" is obtained which accurately reflects the predictive ability of the model.

In addition to q^2, another useful statistical metric is the *t*-value associated with each model descriptor, defined as the descriptor coefficient divided by its standard error. Descriptors with large $|t|$ values are highly significant in the predictive model and, as such, can be examined in order to gain some understanding of the nature of the property or activity of interest. It must be noted, however, that no conclusions may be drawn with respect to descriptors associated with small $|t|$ values.

For the sake of clarity, it should be re-stated that the ridge regression method used in the current study does not involve variable selection, as this is a shrinkage method which is designed to use all available descriptors.

3. Results and Discussion

The major objective of this study was to investigate the utility of graph theoretical invariants in the formulation of QSARs for the anticancer activity of 2-phenylindole derivatives.

Results presented in Table 3 show that, in terms of the predictive power of the models, the TI+AP model ($q^2 = 0.867$) is better than those developed using TI ($q^2 = 0.512$) or AP ($q^2 = 0.653$) alone. The models developed using only topological indices or atom pairs alone are also inferior to that reported by Liao et al. using CoMFA.[12] However, the TI + AP model substantially outperforms the CoMFA model ($q^2 = 0.705$).

Table 3. Ridge regression results with TI, AP, and TI + AP compared with the result from CoMFA analysis.

Descriptor class	q^2	PRESS
Current Study		
TI	0.512	5.976
AP	0.653	12.990
TI+AP	0.867	4.983
CoMFA Result[a]	0.705	[b]

[a] CoMFA result from Liao et al.;[12] [b] PRESS value not available.

Inhibition of microtubule function using tubulin targeting agents is a well established approach to anticancer chemotherapy.[49–53] Over the years, a large number of natural and synthetic small molecules have been identified as colchicine site inhibitors (CSIs) of tubulin. The enormous molecular diversity of the CSIs is of benefit to drug design because it provides a wide variety of molecular scaffolds for optimization. Determining the essential structural features necessary for anticancer activity is, at the same time, a formidable challenge.[54]

Both normal and cancer cells can alter expression of various tubulin isoforms (encoded by different genes) in response to external stimuli that modify microtubule stability. Currently known anti-tubulin drugs bind to all of these isoforms, with a slight preference for one over the others. It is also known that cancer cells express a variety of tubulin isoforms and are not limited to those expressed in the noncancerous cells from which they originate. Therefore, a drug that preferentially binds with a particular isoform present in the cancer cell only could affect those cells selectively, while being relatively non-toxic to normal cells.[55–57]

At the biochemical level, 2-phenylindoles act via perturbation of the colchicine binding sites on tubulin. A common mechanism of action of these compounds is expected from the fact that all 43 compounds analyzed by us and Liao et al.[12] have the same basic structural scaffold. Such structural homogeneity usually helps the alignment process essential for the CoMFA analysis. Yet, it is interesting to note that the QSAR generated in this paper using a diverse set of calculated mathematical descriptors, viz., combination of TIs and APs, significantly outperforms the CoMFA model in terms of predictive power. It is possible that the variety of ligand-biotarget interactions arising from the substitution patterns of the 43 analogs is better represented by the diverse TI + AP set of descriptors as compared to the CoMFA variables.

Table 4 lists the 20 descriptors with highest $|t|$ values for the TI+AP model reported in Table 3. The TIs are classified as either TS or TC. The following classes of molecular descriptors are found to be influential in the QSAR of the 2-phenylindole derivatives:

a) ${}^{6}\chi^{b}_{\mathrm{Ch},}$ ${}^{6}\chi^{v}_{\mathrm{Ch},}$ ${}^{9}\chi_{Ch}$, ${}^{6}\chi_{\mathrm{Ch}}$ which encode information regarding cyclicity of structure of the compounds under investigation.

b) ${}^{6}\chi^{v}_{\mathrm{C}}$ represents the extent of branching in the molecules.

c) $ANV_{1,}$ $ASV_{2,}$ $DSV_{2,}$ $DS1_{1,}$ $DN^{2}1_{1,}$ $DN^{2}N_{2,}$ $AN1_{2,}$ $AS1_{2}$ are triplet indices which characterize the electronic character of the molecules.

d) C1X3_2_N0X2, C1X3_3_N0X2, C1X2_4_C1X2 are atom pairs which represent specific substructures which are influential for ligand-biotarget interaction.

The class of models presented here, viz., RR approach using easily calculated mathematical descriptors

Table 4. Descriptors with largest | t | values taken from the TI + AP model

TI+AP	|t|	Descriptor Class
${}^{6}\chi^{v}_{C}$	29.21	TC
ANV_1	29.19	TS
ASV_2	28.28	TS
DSV_2	28.07	TS
$DS1_1$	28.05	TS
DN^21_1	28.01	TS
${}^{6}\chi^{b}_{Ch}$	27.96	TC
DN^2N_2	27.96	TS
$AN1_2$	27.66	TS
$AS1_2$	27.30	TS
DN^2Z_2	27.30	TC
${}^{6}\chi_{Ch}$	27.29	TS
${}^{6}\chi^{v}_{Ch}$	27.08	TC
$DS1_2$	27.00	TS
DN^21_2	26.85	TS
C1X3_2_N0X2	26.84	AP
C1X3_3_N0X2	26.84	AP
C1X2_4_C1X2	26.84	AP
${}^{6}\chi^{b}_{C}$	26.82	TC
${}^{9}\chi_{Ch}$	*25.49*	*TS*

and a subset of influential descriptors presented in Table 4, can be used in computer-assisted drug design and prediction of toxicological/ ecotoxicological properties of environmental pollutants.

In the area of drug design, since the QSAR model for the phenylindoles was developed based on descriptors which can be calculated fast, the synthetic chemists can use these models as a decision support tool in synthesis planning. For example, in the indole moiety and the other phenyl ring, one can envision a number of sites where substitution of hydrogen by other groups is possible. Hansch and Leo had tabulated a list of 230 substituents for rational drug design.[58] If one wishes to substitute each of R1, R2 and R3 positions of Figure 1 by a small number, say 50, of substituents, the possible number of derivatives will be 50^3 = 125,000. One cannot handle such a large number of chemicals intuitively; but the high quality QSAR of phenylindoles derived in this paper can be used to screen such a large library pretty fast and the compounds which are predicted to be promising by the QSAR model can be synthesized and tested. This line of approach could look like that in Figure 2.

Another way of handling the combinatorial explosion consisting of a virtual library of 125,000 derivatives could be to cluster the large set into a small number, say 50, of clusters using the most important descriptors in Table 4 and select one chemical from each cluster for synthesis and testing. Such a subset of phenylindoles will be structurally diverse and will have the chance of having novel bioactivity profiles. A similar method was used by Lajiness[59] of the Upjohn Company (now part of Pfizer) based on topological indices calculated by the POLLY[28] software to discover quite a few novel drug leads.

In the area of application of RR and topological descriptor based QSARs in the estimation of properties needed by Globally Harmonized System of Classification and Labelling of Chemicals (GHS); Registration, Evaluation, Authorization and Restriction of Chemicals (REACH);

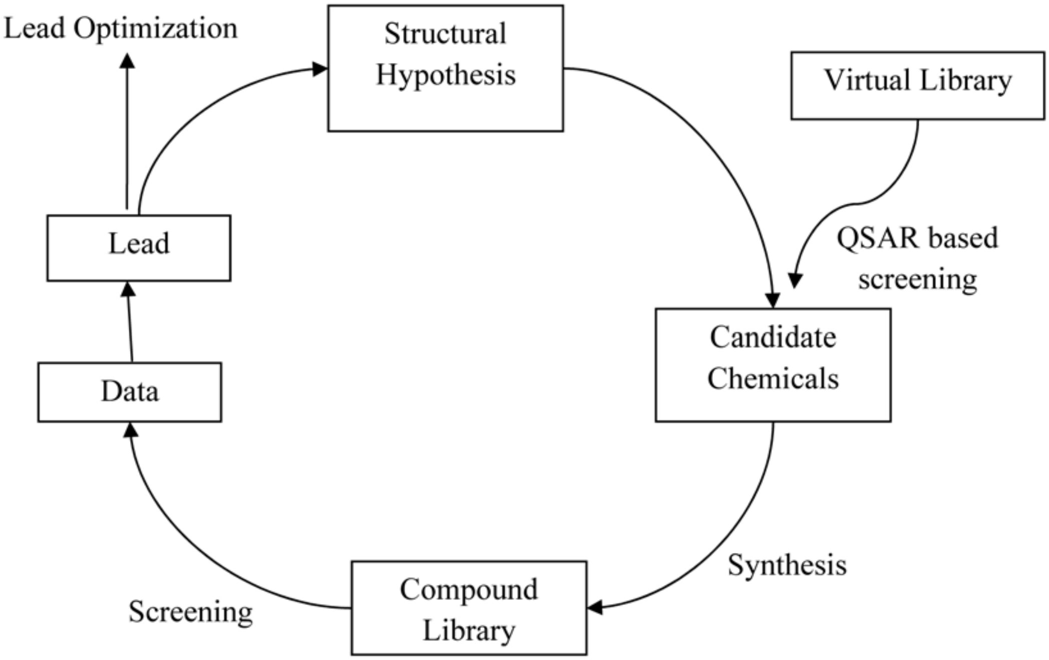

Figure 2. Chemical synthesis assisted by QSAR

and chemical evaluation by agencies like the United States Environmental Protection Agency (USEPA); we can envision a lot of possibility. The GHS needs a large number of health and environmental toxicity data on chemicals, viz., acute toxicity, skin corrosion, skin irritation, eye effects, sensitization, germ cell mutagenicity, carcinogenicity, reproductive toxicity, acute aquatic toxicity, etc.[60] The majority of chemicals currently used in commerce worldwide will *not* have such experimentally determined data sets. For example, the Toxic Substances Control Act (TSCA) Inventory maintained by the USEPA contains more than 83,000 chemicals.[61] Most of these substances do not have experimental physicochemical and toxicological test data prerequisite to their hazard assessment. Therefore, in the foreseeable future property estimation for ecological risk assessment will be carried out on nonempirical ground.[62] Topological descriptors in combination with ridge regression and the hierarchical QSAR (HiQSAR) approach have been useful in the estimation of diverse properties of chemicals like, toxicity and toxic modes of action,[63, 64] vapor pressure,[65] boiling point,[66] dermal penetration,[67] blood: air partition coefficient,[68] Ah receptor binding potency,[69] mutagenicity,[42] allergy contact dermatitis,[70] etc. After the Human Genome Project, a lot of "omics" data are being generated on chemicals of interest. The RR method has been used to combine chemodescriptors and proteomics based biodescriptors in predicting toxicity of priority pollutants like halocarbons.[71] The REACH legislation of the European Community also needs a suite of properties for the evaluation of potential toxicity of new and existing chemicals. For most of the chemicals and their metabolites, such properties are not available. In the area of theoretical descriptor based QSARs, one can use topological indices, substructures, 3-D descriptors or more computationally demanding quantum chemical descriptors. In a series of papers on HiQSARs, we found that for most properties like aryl hydrocarbon receptor binding affinity,[72] mosquito repellency of aminoamides,[18] acute toxicity of benzene derivatives,[73] dermal penetration of polycyclic aromatic hydrocarbons,[74] mutagenicity of aromatic and heteroaromatic amines,[75] mosquito repellency of DEET-related compounds,[76] tissue:air partition coefficients,[21] vapor pressure of 469 diverse compounds,[77] and mutagenicity of 508 diverse compounds,[78] the addiction of quantum chemical indices after the use of topological indices did not improve the predictive power of the models. Therefore, properly validated RR based QSAR models derived from easily calculated descriptors like topological indices and atom pairs as reported in this paper for 2-phenylindoles could be very useful tools for the estimation of various toxicologically and ecotoxicologically relevant properties for hazard assessment of chemicals.

For the proper validation of QSARs needed by regulatory agencies and drug discovery groups for the estimation of potential toxicity of chemicals, the example of RR based QSAR can be applied in many cases. In most practical situations, the number of data points (dependent variables) is small and much smaller than the number of independent variables. Hawkins et al[46, 79] put forward convincing statistical evidence that for small data sets the leave one out method of cross validation is superior to the external validation method. So, it is expected that the type of QSAR exemplified in this paper will have wide applications in drug discovery and hazard assessment of chemicals.

4. Conclusion

Topological indices and atom pairs derived from chemical graph theory produced high-quality models for the prediction of anticancer activity of a set of 43 phenylindole derivatives which act by the disruption of tubulin working through the colchicine binding site. The QSAR formulated using TIs and APs together was superior to the CoMFA model developed from the same set of chemicals. Easily calculated molecular descriptors like TIs and APs used in this paper may find application in the QSAR and *in silico* prediction of bioactivity of potential therapeutic agents in new drug discovery protocols as well as other toxic substances.

5. Acknowledgements

This paper is dedicated to Professor Milan Randic for his outstanding contributions in the fields of mathematical chemistry, development of novel topological descriptors, and QSAR. This is publication # 502 from Center for Water and the Environment, Natural Resources Research Institute, University of Minnesota Duluth, Duluth, MN, USA.

6. References

1. J. R. Williams, C. Shah, D. Sackett, *Anal biochem* **1999**, *275(2)*, 265–267.
2. J. B. Olmsted, G. G. Borisy, *Annu. Rev. Biochem.* **1973**, *42*, 507–540.
3. E. Hamel, *Med. Res. Rev.* **1996**, *16*, 207–231.
4. A. Jordan, J. A. Hadfield, N. J. Lawrence, A. T. McGown, *Med. Res. Rev.* **1998**, *18*, 259–296.
5. Q. Shi, K. Chen, S. L. MorrisNatschke, K. H. Lee, *Curr. Phar. Des.* **1998**, *4*, 219–248.
6. E. Nogales, *Annu. Rev. Biochem.* **2000**, *69*, 277–302.
7. E. Pasquire, N. Andre, D. Braguer, *Curr. Cancer Drug Targets* **2007**, *7*,566–581.
8. K. Odlo, J. Hentzen, J. F. dit Chabert, S. Ducki, O. A. B. S. M. Gani, I. Sylte, M. Skrede, V. A. Florenes, T. V. Hansen, *Bioorg. Med. Chem.* **2008**, *16*, 4829–4838.

9. R. Gastpar, M. Goldbrunner, D. Marko, E. von Angerer, *J. Med. Chem.* **1998**, *41*, 4965–4972.
10. M. Projarova, D. Kaufmann, R. Gastpar, T. Nishino, P. Reszka, P. J. Bednarski, E. von Angerer, *Bioorg. Med. Chem.* **2007**, *15*, 7368–7379.
11. D. Kaufmann, M. Pojarova, S. Vogel, R. Liebl, R. Gastpar, D. Gross, T. Nishino, T. Ptfaller, E. von Angerer, *Bioorg. Med. Chem.* **2007**, *15*, 5122–5136.
12. S. Y. Liao, Q. Li, T. F. Miao, H. L. Lu, K. C. Zheng, *European J. Med. Chem.* **2009**, *44*, 2822–2827.
13. S. C. Basak, B. D. Gute, L. R. Drewes, *Pharm. Res.* **1996**, *13*, 775–778.
14. S. C. Basak and D. Mills, *SAR QSAR Environ. Res.* **2001**, *12*, 481–496.
15. S. C. Basak, D. Mills, B. D. Gute, G. D. Grunwald, and A. T. Balaban, *Topology in Chemistry: Discrete Mathematics of Molecules,* D. H. Rouvray and R. B. King, Eds., Horwood Publishing Limited, Chichester, England, pp. 113–184 (2002).
16. S. C. Basak, B. D. Gute, D. Mills, D. M. Hawkins, *J. Mol. Struct. (Theochem)* **2003**, *622*, 127–145.
17. S. C. Basak, K. Balasubramanian, B. D. Gute, D. Mills, A. Gorczynska, S. Roszak, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1103–1109.
18. S. C. Basak, R. Natarajan, D. Mills, in Conference Proceedings: WSEAS Transactions on Information Science and Applications **2005**, 958–963.
19. S. C. Basak, D. Mills, *ARKIVOC*), **2005,** *10,* 60–76.
20. S. C. Basak, D. Mills, B. D. Gute, R. Natarajan, *Topics in Heterocyclic Chemistry Vol. 5: QSAR and Molecular Modeling Studies of Heterocyclic Drugs,* S. P. Gupta, S. P. Gupta, ed., Springer-Verlag, Berlin-Heidelberg-New York, **2006,** 39–80.
21. S. C. Basak, D. Mills, B. D. Gute, *SAR QSAR Environ. Res.* **2006**, *17*, 515–532.
22. S.C. Basak, D. Mills, and B.D. Gute, *Biological Concepts and Techniques in Toxicology: An Integrated Approach,* J. E. Riviere, ed., Taylor & Francis, New York, **2006,** 61–82.
23. S. C. Basak, D. Mills, *SAR QSAR Environ. Res.* **2009**, *20*, 119–132.
24. S. C. Basak, D. Mills, R. Natarajan, B. D. Gute *Chemical Reactivity Theory: A Density Functional View,* P. K. Chattaraj, ed., CRC Press, Boca Raton, FL, **2009,** 479–502.
25. R. E. Carhart, D. H. Smith, R. Venkataraghavan, *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64–73.
26. S. C. Basak, B. D. Gute, D. Mills, *ARKIVOC* **2006**, *2006*, 157–210.
27. S. C. Basak, G. D. Grunwald, APProbe, Copyright of the University of Minnesota, **1993**.
28. S. C. Basak, D. K. Harriss, V. R. Magnuson, POLLY v. 2.3, Copyright of the University of Minnesota, **1988**.
29. P. A. Filip, T. S. Balaban, A. T. Balaban, *J. Math. Chem.* **1987**, *1*, 61–83.
30. Molconn-Z Version 3.5, Hall Associates Consulting, Quincy, MA, **2000**.
31. L. B. Kier, L. H. Hall, Molecular Connectivity in Structure-Activity Analysis, Research Studies Press, Letchworth, Hertfordshire, U.K., 1986.
32. M. Randic, *J. Am. Chem. Soc.* **1975**, *97*, 6609–6615.
33. A. B. Roy, S. C. Basak, D. K. Harriss, V. R. Magnuson, in Mathl. Modelling Sci. Tech., X.J.R. Avula, R.E. Kalman, A.I. Liapis, E.Y. Rodin, eds., Pergamon Press, New York, **1983**, 745–750.
34. L. B. Kier and L. H. Hall, "Molecular Structure Description: The Electrotopological State", Academic Press, San Diego, CA (1999).
35. SAS Institute, Inc. In SAS/STAT User Guide, Release 6.03 Edition; SAS Institute Inc.: Cary, NC., **1988**.
36. A. E. Hoerl, R. W. Kennard, *Technometrics* **1970**, *12*, 55–67.
37. A. E. Hoerl, R. W. Kennard, *Technometrics* **2005**, *12*, 69–82.
38. I. E. Frank, J. H. Friedman, *Technometrics* **1993**, *35*, 109–135.
39. S. Wold, *Technometrics* **1993**, *35*, 136–139.
40. A. Hoskuldsson, *J. Chemometrics* **1995**, *9*, 91–123.
41. A. Hoskuldsson, *J. Chemometrics* **1988**, *2*, 211–228.
42. S. C. Basak, D. Mills, M. M. Mumtaz, K. Balasubramanian, *Indian J. Chem.* **2003**, *42A*, 1385–1391.
43. S. C. Basak, D. Mills, H. A. El-Masri, M. M. Mumtaz, D. M. Hawkins, *Environ. Toxicol. Pharmacol.* **2004**, *16*, 45–55.
44. S. C. Basak, D. Mills, D. M. Hawkins, H. El-Masri, *Risk Analysis* **2003**, *23*, 1173–1184.
45. S. C. Basak, D. Mills, D. M. Hawkins, H. A. El-Masri, *SAR QSAR Environ. Res.* **2002**, *13*, 649–665.
46. D. M. Hawkins, S. C. Basak, D. Mills, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 579–586.
47. J. J. Kraker, D. M. Hawkins, S. C. Basak, R. Natarajan, D. Mills, Chemometr. *Intell. Lab. Syst.* **2007**, *87*, 33–42.
48. S. C. Basak, R. Natarajan, D. Mills, D. M. Hawkins, J. J. Kraker, *J. Chem. Inf. Model.* **2006**, *46*, 65–77.
49. M. C. Lin, H. H. Ho, G. R. Pettit, E. Hamel, *Biochemistry* **1989**, *28*, 6984–6991.
50. T. Beckers, S. Mahboobi, *Drugs Future* **2003**, *28*, 767–785.
51. Q. Li, H. L.Sham, *Expert Opin. Ther. Pat.* **2002**, *12*, 1663–1702.
52. H. Prinz, Expert Rev. *Anticancer Ther.* **2002**, *2*, 695–708.
53. P. M. Checchi, J. H. Netlles, J. Zhou, P. Snyder, H. C. Joshi, *Trends Pharmacol. Sci.* **2003**, *24*, 361–365.
54. T. L. Nguyen, C. McGrath, A. R. Hermone, J. C. Burnett, D. W. Zaharevitz, B. W. Day, P. Wipf, E. Hamel, R. Gussio, *J. Med. Chem.* **2005**, *48*, 6107–6116.
55. J. Y. Mane, M. Klobukowski, *J. Chem. Inf. Model.* **2008**, *48*, 1824–1832.
56. I. Khan, R. Luduena, *InVest. New Drugs* **2003**, *21*, 3–13.
57. A. Banerjee, R. Luduena, *J. Biol. Chem.* **1992**, *267*, 13335–13339.
58. C. Hansch, A. Leo, Exploring QSAR: Fundamentals and Applications in Chemistry and Biology, American Chemical Society, Washington, DC, **1995**.
59. M. Lajiness, *Computational Chemical Graph Theory*, ed Rouvray DH (Nova, New York), **1990**, 299–316.
60. http://www.osha.gov/dsg/hazcom/ghs.html
61. TSCA Inventory: http://www.epa.gov/lawsregs/laws/tsca.html

62. U. Maran, M. Karelson, A. R. Katritzky, *Quant. Struct.-Act. Relat*., **1999**, *18*, 3–10.
63. G. W. Mushrush, S. C. Basak, J. E. Slone, E. J. Beal, S. Basu, W. M. Stalick, D. R. Hardy, *J. Environ. Sci. Health* **1997**, *A32*, 2201–2211.
64. S. C. Basak, G. D. Grunwald, G. E. Host, G. J. Niemi, S. P. Bradbury, *Environ. Toxicol. Chem.* **1998**, *17*, 1056–1064.
65. S. C. Basak, B. D. Gute, G. D. Grunwald, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 651–655.
66. S. C. Basak, D. Mills, *Commun. Math. Comput. Chem.* **2001**, *44*, 15–30.
67. B. D. Gute, G. D. Grunwald, S. C. Basak, *SAR QSAR Environ.Res.* **1999**, *10*, 1–15.
68. S. C. Basak, D. Mills, D. M. Hawkins, H. A. El-Masri, *Risk Analysis* **2003**, *23*, 1173–1184.
69. D. M. Hawkins, S. C. Basak, D. Mills, *Environ. Toxicol. Pharmacol.* **2004**, *16*, 37–44.
70. S. C. Basak, D. Mills, D. M. Hawkins, *J. Comput. Aided Mol. Des*. **2008**, *22*, 339–343.
71. D. M. Hawkins, S. C. Basak, J. J. Kraker, K. T. Geiss, F. A. Witzmann, *J. Chem. Inf. Model*. **2006**, *46*, 9–16.
72. S. C. Basak, D. Mills, M. M. Mumtaz, K. Balasubramanian, *Indian J. Chem.* **2003**, *42A*, 1385–1391.
73. B. D. Gute, S. C. Basak, *SAR QSAR Environ. Res.* **1997**, *7*, 117–131.
74. B. D. Gute, G. D. Grunwald, S. C. Basak, *SAR QSAR Environ. Res.* **1999**, *10*, 1–15.
75. S. C. Basak, D. R. Mills, A. T. Balaban, B. D. Gute, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 671–678.
76. R. Natarajan, S. C. Basak, D. Mills, J. J. Kraker, D. M. Hawkins, *Croat. Chem. Acta* **2008**, *81(2)*, 333–340.
77. S. C. Basak, D. Mills, *ARKIVOC,* **2005,** *10,* 308-320,
78. S. C. Basak, D. Mills, B. D. Gute, and D. M. Hawkins, *Quantitative Structure-Activity Relationship (QSAR) Models of Mutagens and Carcinogens,* R. Benigni, Ed., CRC Press, Boca Raton, FL, Chapter 7, pp. 207-234 (2003).
79. D. M. Hawkins, J. J. Kraker, S. C. Basak, D. Mills, *SAR QSAR Environ. Res.* **2008**, *19*, 525–539.

Povzetek

Z uporabo topoloških indeksov (TI) in atomskih parov (AP) smo razvili model za kvantitativno določanje odnosa med strukturo in aktivnostjo (QSARs, quantitative structure-activity relationships) za niz 43 derivatov 2-fenilindolov, katerih aktivnost se kaže kot zaviranje rakotvornosti. Rezultati kažejo, da imajo QSAR modeli, osnovani na kombinaciji TI in AP, boljše napovedne zmogljivosti od tistih, ki upoštevajo le TI ali AP. Korelacijski koeficient q^2 modela verižne regresije z uporabo TI + AP je 0.867, v primerjavi z 0.705 iz literaturne študije na osnovi analize komparativnega molekulskega polja (CoMFA).