J Mol Model (2004) 10:102–111
DOI 10.1007/s00894-003-0173-6

ORIGINAL PAPER

Manish C. Bagchi · Bhim C. Maiti · Denise Mills ·
Subhash C. Basak

Usefulness of graphical invariants in quantitative structure–activity correlations of tuberculostatic drugs of the isonicotinic acid hydrazide type

Received: 4 March 2003 / Accepted: 10 November 2003 / Published online: 23 December 2003

Abstract Quantitative structure–activity relationship (QSAR) studies have been performed for a series of 2-substituted isonicotinic acid hydrazides utilizing theoretical molecular descriptors. 223 topological (topostructural and topochemical) indices along with seven geometrical descriptors were computed for the prediction of antibacterial activity against *Mycobacterium tuberculosis*. Ridge-regression models assessed by cross-validated R^2 have been formulated, and a comparative study on the relative effectiveness of physicochemical vis-à-vis theoretical molecular descriptors performed. The models developed clearly indicate the supremacy of structure–activity over property–activity relationships in the current study and can be used to evaluate the potential tuberculostatic activity of other INH derivatives, real or hypothetical.

Keywords Tuberculostatic drugs · Topological indices · Molecular descriptors · Ridge regression · Structure–activity relationships · Physicochemical properties

Introduction

Quantitative structure–activity relationship (QSAR) studies are based on the premise that biological response is a function of chemical structure. Thus, significant parameters of chemical structure have been defined in numerical terms for use in the development of specific QSAR models. [1] This paradigm leads to the belief that a proper choice of chemical structural descriptors will give a reasonable prediction of biological response for molecules. A recent interest in pharmaceutical drug design and hazard assessment of chemicals is the prediction of environmental, physicochemical, toxicological and pharmacological properties of chemicals directly from their structure. [2] Early QSAR studies used physical properties and physicochemical substituent constants for the prediction of other more complex physicochemical, biomedicinal, and toxicological properties. Such property–property correlations are useful only when properties necessary for prediction are available for all chemicals under consideration. In contemporary drug design, one can produce large real or virtual combinatorial libraries of chemicals for screening. Most of these chemicals have no physicochemical data, and thus predictive methods based on experimental data are of limited use in this situation. Hence, there is a need for the development of QSAR methods using non-empirical parameters. A recent trend in this direction is the use of theoretical molecular descriptors, which can be calculated directly from molecular structure. Topological indices or numerical graph invariants constitute an important subset of these theoretical descriptors. Topological indices are derived from different classes of weighted graphs, representing various levels of chemical structural information. They are numerical quantifiers of molecular topology and encode information regarding size, shape, branching pattern, cyclicity, and symmetry of molecular graphs. Topostructural, topochemical, and geometrical (3D) indices have been widely used in QSAR research for predicting biological activities in rational drug design. A large number of QSARs pertaining to chemistry, pharmacology, and toxicology have used these non-empirical parameters [3, 4, 5] in the form of mathematical models that relate molecular structure to their physicochemical, biomedicinal, and toxic properties.

M. C. Bagchi (✉)
Drug Design, Development and Molecular Modelling Division,
Indian Institute of Chemical Biology,
4 Raja S.C. Mullick Road, Jadavpur, Calcutta 700032, India
e-mail: mcbagchi@iicb.res.in
Tel.: +91 33 2473 3491/3493/0493/6793
Fax: +91 33 2473 5197

B. C. Maiti
Chemistry of Bioactive Substances Division,
Indian Institute of Chemical Biology,
4 Raja S. C. Mullick Road, Jadavpur, Calcutta 700032, India

D. Mills · S. C. Basak
Natural Resources Research Institute,
University of Minnesota-Duluth,
5013 Miller Trunk Highway, Duluth, MN 55811, USA

The present paper aims at developing QSARs for tuberculostatic drugs and their analogs using topostructural, topochemical, and geometrical (3D) indices. Seydel et al. [6] formulated QSARs for the same set of INH derivatives based on physicochemical descriptors viz., π, p$K_{a,}$ $E_{s,}$ and $V_{w.}$ Such models will be of limited utility in the evaluation of potential tuberculostatic activity of a larger and more structurally diverse group of INH derivatives because properties such as pK_a and E_s for most of those chemicals will be unavailable. A viable alternative under such circumstances is the development of QSARs using theoretical molecular descriptors. To this end, we have carried out a comparative study of the relative effectiveness of physicochemical vis-à-vis calculated molecular descriptors, viz., topostructural, topochemical, and geometrical parameters, in the QSAR of INH derivatives.

The results are presented here along with the utility and limitations of the QSAR models.

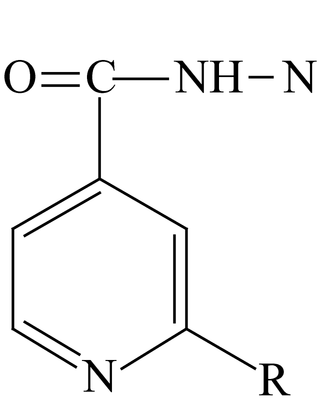

Fig. 1 2-Substituted INH

Methods

Biological activity data of isoniazide

The action of isonicotinic acid hydrazide against *Mycobacterium tuberculosis* has been studied by Seydel et al. [6] considering 2-substituted INH derivatives (see Fig. 1). They synthesized 19 such derivatives in order to study the electronic, steric, and hydrophobic properties of the substituents. The biological activity data in the form of minimum inhibitory concentration (MIC in μM) were determined experimentally (Table 1). They developed QSAR models for these 2-substituted INH derivatives using mainly a few physicochemical parameters such as steric effect, electronic effect, van der Waals' volume, and basicity. The number of available physicochemical parameters is limited. On the other hand, a much larger number of theoretical molecular descriptors is available to define the structural variety of a set of molecules explicitly. So, these may be considered for the construction of a valid QSAR model. QSAR models developed by using experimental properties as independent variables are essentially property–property correlations, whereas models developed using descriptors based solely on molecular structure throw light on structure–property correlations, which may provide a better tool for rational drug design [7, 8, 9].

Theoretical molecular descriptors

The molecular descriptors used in this study are of three categories—(a) topostructural (TS), (b) topochemical (TC), and (c) geometrical (3D). Topostructural descriptors encode information strictly on the neighborhood and connectivity of atoms within the molecule, while the topochemical descriptors encode information related to both the topology of the molecule and the chemical nature of atoms and bonds within it. The three-dimensional or shape descriptors (3D) are more complex, encoding information about the three-dimensional aspects of molecular structure. With the hierarchical QSAR method, multiple models are developed, each time including an additional descriptor class that is more complex and computationally demanding. Comparing the statistical metrics of the hierarchically developed models, the relative contribution of each descriptor class can be examined.

In our present study, the software packages, POLLY, [10] Triplet, [11, 12] and Molconn-Z, [13] have been used for the calculation of molecular descriptors. From POLLY and associated software, a set of 102 topological descriptors is available, including a large group of connectivity indices and path-length descriptors, [14, 15, 16, 17] Balaban's J indices, [18, 19, 20] and information theoretic descriptors including neighborhood complexity indices. [21, 22] The Triplet program calculates a set of 100 topological

Table 1 Substituents and properties of 2-substituted INH derivatives. Data obtained from Seydel et al. [6]

Compd	R	MIC	Log 1/MIC	π	pK_a	V_w
1	H	1.1	−0.041	0	5.17	3.45
2	CH_3	5.2	−0.716	0.769	5.94	13.67
3	C_2H_5	21.1	−1.324	1.253	5.97	23.9
4	n-C_3H_7	55.2	−1.742	1.765	5.97	34.13
5	i-C_4H_9	450.0	−2.653	2.162	5.97	44.35
6	CH_3O	153.0	−2.185	1.04	3.06	16.87
7	C_2H_5O	450.0	−2.655	1.62	3.47	27.1
8	NH_2	14.5	−1.161	0.16	6.71	10.54
9	CH_3CONH	2150.0	−3.332	−0.11	4.09	33.45
10	$CH_3CONHCH_2$	243.0	−2.386	−0.439	4.23	43.68
11	$(C_2H_5)_2N$	717.0	−2.856	2.254	7.32	52.13
12	F	260.0	−2.415	1.03	−0.44	5.8
13	Cl	392.0	−2.593	1.25	0.72	12
14	Br	616.0	−2.79	1.39	0.9	15.12
15	I	254.0	−2.404	1.652	1.82	19.64
16	NO_2	371.0	−2.569	0.378	−2.2	16.8
17	C_6H_5	50.0	−1.699	2.492	4.48	45.84
18	$C_6H_5CH_2$	38.5	−1.585	2.145	5.13	56.07
19	CH_2=CH	35.0	−1.544	1.53	4.98	20.41

Table 2 Symbols, definitions and classification of calculated molecular descriptors

Topostructural (TS)	
I^{W}_{D}	Information index for the magnitudes of distances between all possible pairs of vertices of a graph
$\overline{I^{W}_{D}}$	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
$\overline{H^{D}}$	Graph distance complexity
$\overline{\mathrm{IC}}$	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
${}^{h}\chi_{\mathrm{C}}$	Cluster connectivity index of order h=3–6
${}^{h}\chi_{\mathrm{PC}}$	Path-cluster connectivity index of order h=4–6
${}^{h}\chi_{\mathrm{Ch}}$	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
$\mathrm{DN}^2\mathrm{S}_y$	Triplet index from distance matrix, square of graph order (# of non-H atoms), and distance sum; operation y=1–5
$\mathrm{DN}^2 1_y$	Triplet index from distance matrix, square of graph order, and number 1; operation y=1–5
$\mathrm{AS1}_y$	Triplet index from adjacency matrix, distance sum, and number 1; operation y=1–5
$\mathrm{DS1}_y$	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
DSN_y	Triplet index from distance matrix, distance sum, and graph order; operation y=1–5
$\mathrm{DN}^2\mathrm{N}_y$	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
$\mathrm{AN1}_y$	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
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
${}^{h}\chi^{\mathrm{b}}$	Bond path connectivity index of order h=0–6
${}^{h}\chi^{\mathrm{b}}_{\mathrm{C}}$	Bond cluster connectivity index of order h=3–6
${}^{h}\chi^{\mathrm{b}}_{\mathrm{Ch}}$	Bond chain connectivity index of order h=3–6
${}^{h}\chi^{\mathrm{b}}_{\mathrm{PC}}$	Bond path-cluster connectivity index of order h=4–6
${}^{h}\chi^{\mathrm{v}}$	Valence path connectivity index of order h=0–6
${}^{h}\chi^{\mathrm{v}}_{\mathrm{C}}$	Valence cluster connectivity index of order h=3–6
${}^{h}\chi^{\mathrm{v}}_{\mathrm{Ch}}$	Valence chain connectivity index of order h=3–6
${}^{h}\chi^{\mathrm{v}}_{\mathrm{PC}}$	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
HB_1	Hydrogen bonding parameter
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
$\mathrm{DN}^2\mathrm{Z}_y$	Triplet index from distance matrix, square of graph order, and atomic number; operation y=1–5
nvx	Number of non-hydrogen atoms in a molecule
nelem	Number of elements in a molecule
fw	Molecular weight
${}^{h}\chi^{\mathrm{v}}$	Valence path connectivity index of order h=7–10
${}^{h}\chi^{\mathrm{v}}_{\mathrm{Ch}}$	Valence chain connectivity index of order h=7–10
si	Shannon information index

Table 2 (continued)

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 (kp1$*$kp2/nvx)
IdCbar	Bonchev–Trinajstic information index
IdC	Bonchev–Trinajstic 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
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, –NH2, –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, –NH2, –NH–, >N–, –O–, –S–, along with –F and –Cl
Qv	General Polarity descriptor
NHBint$_y$	Count of potential internal hydrogen bonders (y=2–10)
SHBint$_y$	E-State descriptors of potential internal hydrogen bond strength (y=2–10)
	Electrotopological State index values for atoms 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, SsssssP, SsSH, SdS, SssS, SaaS, SdssS, SddssS, SssssssS, SsCl, SsGeH3, SssGeH2, SsssGeH, SssssGe, SsAsH2, SssAsH, SsssAs, SdsssAs, SsssssAs, SsSeH, SdSe, SssSe, SaaSe, SdssSe, SddssSe, SsBr, SsSnH3, SssSnH2, SsssSnH, SssssSn, SsI, SsPbH3, SssPbH2, SsssPbH, SssssPb
Geometrical (3D)	
kp0	Kappa zero
kp1–kp3	Kappa simple indices
ka1–ka3	Kappa alpha indices

parameters. They are derived from a matrix, a main diagonal column vector, and a free-term column vector, converting the matrix into a system of linear equations whose solutions are the local vertex invariants. These local vertex invariants are then used in various mathematical operations in order to obtain the triplet descriptors. From the Molconn-Z program, we obtain 167 additional descriptors including an extended set of connectivity indices, electrotopological indices [23, 24] and hydrogen bonding descriptors, along with molecular-shape descriptors. A brief description of the set of theoretical molecular descriptors calculated for use in the present study is provided in Table 2.

Statistical analysis

Prior to model development, the set of calculated descriptors was reduced from 369 to 230. The descriptors eliminated include those with a constant value for all, or nearly all, of the compounds, and those that were perfectly correlated (r=1.0) with another descriptor according to the CORR procedure of the SAS statistical package. [25] In addition, the 230 descriptors were transformed by the natural logarithm due to the fact that their scales differed by several orders of magnitude.

Conventional regression (ordinary least squares, OLS) does not produce reliable models when the number of descriptors exceeds the number of observations. [26, 27] In this situation, appropriate statistical methods include ridge regression (RR), [28] principal components regression (PCR), [29] and partial least squares (PLS). [30, 31, 32] Each of these methods is useful when the number of independent variables greatly exceeds the number of observations and when the independent variables are highly intercorrelated. Each of these methods makes use of the entire available pool of independent variables as opposed to selecting a subset, which introduces bias and may result in the elimination of important parameters from the study. Formal comparisons have consistently shown subsetting to be less effective than alternative methods, such as these, that retain all of the independent variables and use other approaches to deal with the rank deficiency. [26, 33] Statistical theory suggests that RR is the best of the three methods, and this has been generally borne out in multiple comparative studies. [9, 33, 34, 35] For this reason, the models based on the large set of TS, TC, and 3D theoretical descriptors were developed using the RR methodology. 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 "shrinks" them differentially according to their eigenvalues. The RR vector of regression coefficients, **b**, is given by

$$\mathbf{b} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + k\mathbf{I}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$$

where **X** is the matrix of descriptors, **Y** is the vector of observed activities, **I** is an identity matrix, and k is a non-negative constant known as the "ridge" constant. [36] If k=0, RR reduces to conventional OLS regression.

Calculations were performed using a Fortran 95 code implementing the "faster ridge" regression algorithm. [37] Cross-validation is used to select the value of *k*. [38] It is important to note that standard regression measures including R^2 are meaningless in the assessment of models based on a large number of descriptors with respect to the number of observations. The value of R^2 tends to increase upon the addition of any descriptor, even those that are irrelevant, possibly resulting in an overestimation of model quality. For that reason, we have reported the cross-validated R^2, which, unlike R^2, tends to *decrease* upon the addition of irrelevant descriptors, providing a reliable measure of model quality. While R^2 is necessarily a positive value, the cross-validated R^2 may be negative, indicating that the associated model is very poor. The cross-validated R^2 is calculated using the leave-one-out approach, wherein each compound is removed, in turn, from the data set and the regression is fitted based on the remaining $n-1$ compounds. The cross-validated R^2 mimics the results of applying the final regression to a future compound; large values can be interpreted unequivocally and without regard to the number of compounds or descriptors as indicating that the model will accurately predict the activity of a compound of the same chemical type as those used to calibrate the regression. The cross-validated R^2 is defined by:

$$R^2_{\text{cv}} = 1 - \frac{PRESS}{SSTotal}$$

where *SSTotal* is the total sum of squares and *PRESS* is the prediction sum of squares, i.e., the sum of squares of the difference between the actual observed activity and that predicted from the regression. As it is based on compounds that are external to the fitted regression, similar to using a test set, it is a reliable measure of model predictability. When the available sample size is small, the leave-one-out cross-validation approach is preferred over holding back a portion for testing. [38]

The set of 230 calculated molecular descriptors was partitioned into TS, TC, and 3D classes, and the RR models were developed utilizing these classes in a hierarchical fashion. In addition to providing values for both the ridge constant and cross-validated R^2, the RR code also provides the *t* value for each descriptor, which is the coefficient estimate divided by its standard error. The |*t*| values can be examined to identify descriptors that are significant for the prediction of antibacterial activity against *Mycobacterium tuberculosis*. While large a value indicates that the associated descriptor is important in the ridge regression model, it is important to note that the converse is not necessarily true. The ridge regression method was also utilized to analyze the data obtained by Seydel et al. [6] in which only three independent variables were used, viz., π, $pK_{a,}$ $V_{w.}$

Table 3 Regression summary for QSARs/QSPRs of INH derivatives

Descriptors	R^2_{cv}	Ridge constant (*k*)
N=19		
Computed molecular descriptors		
TS	0.308	4.0082
TS+TC	0.783	0.0773
TS+TC+3D	0.785	0.0100
TC	0.776	0.6282
3D	0.213	0.0100
Physicochemical descriptors[a]		
$pK_a+\pi$	–0.044	14.567
N=15 (compounds 8–11 omitted)		
Computed molecular descriptors		
TS	–0.044	7.8963
TS+TC	0.736	0.0100
TS+TC+3D	0.737	0.0100
TC	0.781	0.0100
3D	–0.110	0.6551
Physicochemical descriptors[a]		
$pK_{a+}\pi$	0.547	–0.0100
N=15 (compounds 10, and 17–19 omitted)		
Computed molecular descriptors		
TS	0.217	9.6248
TS+TC	0.851	0.0536
TS+TC+3D	0.852	0.0680
TC	0.853	1.1472
3D	0.386	0.1247
Physicochemical descriptors[a]		
pK_a+V_w	0.643	0.9443
N=15 (compounds 5, 9,10 and 16 omitted)		
Computed molecular descriptors		
TS	0.005	6.4540
TS+TC	0.901	0.0100
TS+TC+3D	0.899	0.0100
TC	0.915	0.0100
3D	–0.001	0.7750
Physicochemical descriptors[a]		
pK_a+V_w	0.183	2.5077

[a] Physicochemical data obtained from Seydel et al. [6]

Results and discussion

QSAR studies were performed using theoretical molecular descriptors and experimental biological activity data for 2-substituted INH derivatives. Models were developed for the complete set of 19 such compounds as well as for various subsets based on the work of Seydel et al. [6] The models developed by that research group utilize physicochemical properties including π and pK_a.

Table 3 provides regression results for both the hierarchical RR studies utilizing the calculated set of theoretical descriptors and the RR studies based on physicochemical properties obtained by Seydel et al. [6] For the complete set of 19 compounds, the RR model utilizing the TS+TC descriptors has an R^2_{cv} value of 0.783. The addition of the 3D descriptors does not result in significant model improvement. The TS or 3D descriptors alone result in inferior models. For the same set of compounds, the RR model based on π and pK_a is very poor with an R^2_{cv} value of –0.044. The RR model utilizing TS+TC descriptors for the complete set of 19 compounds can be found in Table 4. Note that the descriptors are sorted by |*t*| values.

When we consider a group of 15 compounds excluding **8**–**11** since all of them possess the amino function that has a great influence on the basicity of the pyridine nitrogen atom, the ridge-regression model based on π and pK_a improves significantly with an R^2_{cv} value of 0.547. However, a superior model is obtained using the TC descriptors, with an R^2_{cv} value of 0.781. Again, the TS and 3D descriptors produce poor models.

The reason for excluding compounds **10** and **17**–**19** from our next analysis lies in the fact that compounds **17** and **19** possess steric effect of the coplanar group (phenyl and vinyl) over the ring nitrogen atom, and compounds **10**

Table 4 TS+TC ridge regression model for the prediction of antibacterial activity against *Mycobacterium tuberculosis* for 19 compounds (descriptors sorted by |*t*|, where *t*=the coefficient divided by its standard error)

Descriptor[a]	RR coeff	s.e.	*t*
CONSTANT	62.57233		
ASZ_3	−1.70146	0.02759	−61.67
DN^2Z_3	−3.04134	0.05315	−57.22
phia	−0.32877	0.00703	−46.76
ANZ_3	−0.51667	0.01154	−44.79
DN^21_1	1.4449	0.03407	42.41
DN^2N_2	1.78358	0.04237	42.09
H^D	−0.18771	0.00459	−40.93
$AN1_5$	−1.70773	0.04321	−39.52
$AN1_1$	−1.6685	0.04356	−38.3
$\bar{I}_D^W$	−0.14035	0.00372	−37.75
$^9\chi$	0.28405	0.00779	36.44
AZV_4	−0.17063	0.00474	−36
J	−0.6745	0.01888	−35.73
$AN1_2$	1.28988	0.03614	35.7
SHCsatu	0.43671	0.01248	35
ASZ_5	−0.17325	0.005	−34.62
IdCbar	−0.20484	0.00628	−32.6
Hmax	−11.50682	0.36801	−31.27
$AS1_2$	7.56902	0.25196	30.04
ANZ_1	−0.26631	0.00888	−29.97
DN^2S_4	−0.02879	0.00097	−29.61
IdC	−0.01729	0.00059	−29.34
$DS1_1$	0.4294	0.01499	28.64
$^0\chi^v$	−0.15871	0.00565	−28.11
ASN_3	−0.07865	0.0028	−28.09
$DS1_2$	19.6845	0.70437	27.95
$^0\chi^b$	−0.14918	0.00534	−27.91
IC_0	−0.99763	0.03613	−27.61
DN^21_2	194.53393	7.20562	27
I^D	−0.15209	0.00572	−26.6
ASV_2	1.57975	0.05977	26.43
nelem	−1.18055	0.04525	−26.09
$DS1_5$	0.66884	0.02618	25.55
$AS1_5$	0.2731	0.01075	25.4
SdO	0.18624	0.00742	25.09
SHCsats	−0.30328	0.01212	−25.03
asz1	−0.11207	0.00456	−24.55
$^9\chi^v$	2.2061	0.09047	24.38
$^6\chi^v_{Ch}$	2.85541	0.11739	24.32
$SHBint_2$	−1.27723	0.05259	−24.28
CIC_1	0.11645	0.00482	24.18
$SHBint_3$	−1.10891	0.04599	−24.11
J^Y	−0.47523	0.01972	−24.1
SIC_0	−0.91954	0.03866	−23.79
$NHBint_7$	0.16761	0.00737	22.75
J^B	−0.47693	0.02098	−22.73
DSV_2	2.597	0.11604	22.38
$^6\chi^b_{Ch}$	2.84585	0.12815	22.21
$AS1_1$	0.24326	0.01097	22.17
fw	−0.16828	0.0076	−22.15
$^3\chi^b_C$	−0.7275	0.03287	−22.13
AZV_5	0.11103	0.00517	21.46
ANZ_5	−2.87527	0.13713	−20.97
$^5\chi^b_{PC}$	−0.35103	0.01679	−20.91
$^2\chi^b$	−0.19433	0.00934	−20.81
ASN_1	−0.57918	0.02787	−20.78
$^5\chi_C$	1.95727	0.09457	20.7
$^{10}\chi$	0.51422	0.02484	20.7
$NHBint_5$	−0.27741	0.01367	−20.29
SdssC	0.23335	0.01176	19.85
$AS1_3$	−1.33631	0.06866	−19.46
SIC_1	−0.50634	0.02603	−19.45
P_6	−0.07043	0.00362	−19.45
DSN_3	−0.07779	0.00406	−19.16
DSN_1	−1.36341	0.07234	−18.85
$^{10}\chi^v$	4.38955	0.23458	18.71
CIC_0	0.2481	0.01344	18.46
DN^2Z_5	−0.25233	0.01376	−18.34
$^6\chi_{Ch}$	0.94722	0.05294	17.89

Table 4 (continued)

Descriptor[a]	RR coeff	s.e.	*t*
$\mathrm{I}^{\mathrm{W}}{}_{\mathrm{D}}$	−0.01014	0.00057	−17.82
ASN_5	−0.52222	0.02936	−17.79
SaaCH	0.08739	0.00491	17.79
$^3\chi^{\mathrm{v}}{}_{\mathrm{C}}$	−0.74437	0.04192	−17.76
P_7	−0.05659	0.00325	−17.43
$^1\chi^{\mathrm{b}}$	−0.19129	0.0111	−17.23
ANN_2	−0.03178	0.00186	−17.07
DSZ_1	−0.10793	0.00638	−16.92
Wt	0.00902	0.00053	16.88
$^4\chi_{\mathrm{PC}}$	0.33786	0.02004	16.86
$^6\chi^{\mathrm{b}}$	−0.33737	0.02012	−16.77
ANN_1	−0.03133	0.00194	−16.19
SsNH2	0.27825	0.01729	16.09
ANS_2	−0.01296	0.00082	−15.86
$^6\chi$	−0.13491	0.00853	−15.81
$^5\chi^{\mathrm{b}}{}_{\mathrm{C}}$	5.46505	0.34602	15.79
DSV_1	0.18483	0.0117	15.79
ANV_2	0.84678	0.05481	15.45
ANN_3	−0.03018	0.00202	−14.94
$\mathrm{DN}^2\mathrm{Z}_1$	−0.13794	0.00924	−14.93
$\mathrm{AN1}_3$	−0.07029	0.00473	−14.86
SHsNH2	0.34602	0.0233	14.85
H^{V}	−0.33398	0.02263	−14.76
$^0\chi$	−0.04998	0.00343	−14.56
P_8	−0.02136	0.00147	−14.51
I_{ORB}	−0.6977	0.04884	−14.29
SaaN	0.36378	0.02557	14.23
$^7\chi$	−0.15035	0.01065	−14.12
P_4	−0.05531	0.00394	−14.05
W	−0.01013	0.00072	−13.99
AZN_5	0.01638	0.00119	13.82
O_{ORB}	0.13141	0.00961	13.68
ANS_1	−0.01593	0.00117	−13.56
SaasC	0.10506	0.00792	13.26
ANZ_2	−0.03483	0.00266	−13.11
P_0	−0.02808	0.00215	−13.06
J^{X}	−0.39307	0.03029	−12.98
ANS_3	−0.01964	0.00152	−12.96
ANS_5	−0.01482	0.00114	−12.96
AZV_3	0.02918	0.00228	12.82
ASV_1	0.15495	0.01227	12.62
$\mathrm{DN}^2\mathrm{1}_4$	−0.00662	0.00054	−12.26
$\mathrm{DN}^2\mathrm{N}_3$	−0.06184	0.00513	−12.04
ANV_4	−0.01253	0.00104	−12
ASV_4	−0.00869	0.00075	−11.65
CIC_3	0.09915	0.00889	11.16
$^5\chi_{\mathrm{PC}}$	−0.18333	0.0165	−11.11
tets2	0.07728	0.00701	11.03
$^1\chi$	−0.0287	0.00266	−10.77
sumDELI	−0.09057	0.00861	−10.52
numHBd	0.23755	0.02263	10.5
AZV_1	0.04571	0.00437	10.47
AZN_4	−0.65722	0.06277	−10.47
$^5\chi^{\mathrm{v}}{}_{\mathrm{C}}$	6.34556	0.61726	10.28
asv5	0.23641	0.02337	10.12
ANZ_4	0.08853	0.00887	9.98
$\mathrm{DN}^2\mathrm{Z}_2$	0.29919	0.03024	9.89
$^2\chi^{\mathrm{v}}$	−0.14645	0.0151	−9.7
$\mathrm{DS1}_3$	−1.91395	0.19797	−9.67
knotpv	0.11053	0.01147	9.63
ANS_4	0.72127	0.07555	9.55
$\mathrm{DN}^2\mathrm{N}_4$	−0.00731	0.00077	−9.45
SHarom	0.06979	0.00759	9.19
NHBint_6	0.11886	0.01307	9.09
ANV_3	−0.027	0.00298	−9.08
CIC_4	0.0817	0.00917	8.91
$^4\chi^{\mathrm{b}}$	−0.14635	0.01701	−8.6
Qv	−0.17802	0.02076	−8.58
numHBa	−0.30769	0.03601	−8.55
IC_1	−1.15432	0.14277	−8.09

Table 4 (continued)

Descriptor[a]	RR coeff	s.e.	*t*
SIC_3	−0.56238	0.07172	−7.84
P_{10}	0.02857	0.00375	7.62
AZN_1	0.01019	0.00136	7.47
${}^5\chi$	0.04382	0.00589	7.44
$\mathrm{DN}^2\mathrm{S}_3$	−0.02209	0.00308	−7.17
ANV_1	0.19003	0.0265	7.17
${}^5\chi^{\mathrm{v}}{}_{\mathrm{PC}}$	−0.09574	0.01377	−6.95
M_2	0.009	0.00134	6.7
${}^3\chi$	0.09536	0.0144	6.62
dn2n1	−3.69449	0.57442	−6.43
P_2	0.01748	0.00279	6.26
SHHBd	0.11932	0.01907	6.26
$\mathrm{AN1}_4$	−0.00556	0.00092	−6.05
AZN_2	0.00608	0.00101	6.04
K_9	−0.00722	0.00122	−5.93
ANN_4	0.01386	0.00236	5.89
SIC_4	−0.43608	0.075	−5.81
ASV_3	−0.3262	0.05611	−5.81
${}^4\chi$	−0.06879	0.01206	−5.7
totop	0.04427	0.008	5.53
$\mathrm{DN}^2\mathrm{1}_3$	−5.97255	1.09566	−5.45
AZS_4	0.02874	0.0053	5.42
AZS_5	0.00413	0.00077	5.37
ASZ_4	0.00623	0.00116	5.34
Hmin	−0.12756	0.02506	−5.09
SHHBa	−0.07508	0.01475	−5.09
knotp	0.08654	0.01719	5.03
${}^6\chi^{\mathrm{v}}$	−0.22586	0.04591	−4.92
Gmin	−0.04729	0.01031	−4.59
${}^6\chi^{\mathrm{b}}{}_{\mathrm{PC}}$	−0.06618	0.01456	−4.54
ASN_2	−0.07281	0.01618	−4.5
${}^7\chi^{\mathrm{v}}$	−0.34436	0.07668	−4.49
P_5	−0.01735	0.0039	−4.45
$\mathrm{AS1}_4$	−0.00314	0.00071	−4.44
$\mathrm{DN}^2\mathrm{N}_5$	−2.56211	0.58342	−4.39
AZN_3	0.00714	0.00163	4.38
DSZ_2	0.027	0.00627	4.31
${}^6\chi^{\mathrm{v}}{}_{\mathrm{PC}}$	0.03996	0.00984	4.06
${}^3\chi^{\mathrm{v}}$	0.07582	0.01879	4.03
${}^6\chi_{\mathrm{PC}}$	0.06054	0.01635	3.7
SHother	0.04615	0.01278	3.61
${}^1\chi^{\mathrm{v}}$	−0.08899	0.02604	−3.42
AZS_1	0.0031	0.00094	3.3
M_1	0.00619	0.00192	3.23
CIC_2	0.13041	0.04352	3
SHssNH	−0.04221	0.01411	−2.99
nclass	−0.07206	0.02575	−2.8
$\mathrm{DN}^2\mathrm{S}_1$	−0.01315	0.00488	−2.69
si	−0.36808	0.1416	−2.6
DSN_4	0.00644	0.0025	2.58
SIC_2	−0.7989	0.31456	−2.54
SssNH	0.03072	0.01215	2.53
${}^2\chi$	0.02708	0.01092	2.48
${}^5\chi^{\mathrm{b}}$	−0.04107	0.01669	−2.46
P_1	−0.00335	0.00141	−2.38
${}^8\chi^{\mathrm{v}}$	0.18971	0.0854	2.22
P_3	0.01001	0.00461	2.17
$\mathrm{DN}^2\mathrm{S}_2$	0.03625	0.01794	2.02
ASN_4	0.00235	0.00125	1.88
ASZ_2	0.00799	0.00459	1.74
IC_4	0.07196	0.04697	1.53
${}^4\chi^{\mathrm{b}}{}_{\mathrm{PC}}$	0.07463	0.04899	1.52
sumI	−0.01161	0.00832	−1.4
AZS_3	0.00175	0.00126	1.39
ANV_5	0.06262	0.05194	1.21
IC_2	−0.70373	0.60151	−1.17
IC_3	−0.05953	0.05412	−1.1
${}^8\chi$	0.01116	0.01111	1
${}^5\chi^{\mathrm{v}}$	0.02774	0.02848	0.97
${}^4\chi^{\mathrm{v}}{}_{\mathrm{PC}}$	0.04066	0.04273	0.95

Table 4 (continued)

Descriptor[a]	RR coeff	s.e.	*t*
DN^2Z_4	−0.00088	0.00098	−0.9
DSN_5	−0.12912	0.16501	−0.78
AZS_2	0.00048	0.00064	0.74
DSN_2	−0.01294	0.01908	−0.68
DN^2S_5	−0.00334	0.00516	−0.65
$\overline{IC}$	−0.07135	0.11881	−0.6
$^3\chi^b$	0.01043	0.01951	0.53
AZV_2	0.00999	0.02312	0.43
SssCH2	0.00574	0.0141	0.41
$^4\chi^v$	0.00418	0.01326	0.32
$DS1_4$	−0.0004	0.00122	−0.32
SsCH3	0.00066	0.00519	0.13
Gmax	0.02201	0.18553	0.12
$^3\chi_C$	−0.00202	0.02576	−0.08

[a] Brief descriptions are provided in Table 2

and **18** possess anomalous values for van der Waals' volume. It is worthwhile to mention that the van der Waals' volume for compounds **10** and **18** are not available in the literature and the values calculated on the basis of adding fragmented V_w (C_6H_5) and V_w (CH_2) were much larger than expected. These values were arbitrarily corrected by Bondi as referred in [6], to establish linearity with other substituents. It can be seen from the result when such a group of 15 INH derivatives was taken into account, the R^2_{cv} in the ridge-regression model utilizing the TC indices alone yields a value of 0.853, whereas the RR analysis based on pK_a and V_w is associated with a value of 0.643 for the same metric.

The last subset of 15 compounds examined in this study was derived by omitting four compounds that were found to be highly influential upon the RR model. The RR model developed utilizing the TC descriptors alone results in an R^2_{cv} of 0.915. Although Seydel et al. did not provide a model for this subset of compounds, we find an R^2_{cv} value of 0.183 upon ridge regression analysis utilizing pK_a and V_w as independent variables. The strong influence of INH derivatives **5**, **9**, **10** and **16** on QSAR models can be discussed in terms of the hypothesis developed by Kruger-Thiemer. According to this hypothesis, ready quaternization of the pyridine nitrogen atom of isonicotinic acid (INA) derivatives in the bacterial cell is essential for its antibacterial activity. The steric effect of the bulky *i*-C_4H_9 group in compound **5**, as well as the bulky amino derivatives in compounds **9** and **10**, decreases the basic character of the pyridine nitrogen atom thus lowering the antibacterial effect considerably. In compound **16**, both the steric as well as the electron withdrawing effect of the polar and planar nitro group causes a considerable decrease in basic character thereby decreasing considerably the R^2_{cv} value when this compound was included in the QSAR analysis.

It is evident from the QSARs reported in Table 3 that the topochemical indices alone can provide a good quality predictive model for 2-substituted INH derivatives. Comparatively, the QSPR studies utilizing the small set of physicochemical properties as molecular descriptors resulted in much inferior models. QSAR models based on purely calculated structural descriptors reported in this paper can be used in evaluating the tuberculostatic potential of any INH derivative, real or hypothetical, and can thus pave the way for the design of novel tuberculostatic drugs.

Acknowledgements Work carried out at the Natural Resources Research Institute and reported in this paper was supported, in part, by Grant F49620-02-1-0138 from the United States Air Force and Cooperative Agreement Number 572112 from the Agency for Toxic Substances and Disease Registry (ATSDR). This is contribution number 342 from the NRRI Center for Water and the Environment. The authors are thankful to Dr. Douglas M. Hawkins of the School of Statistics, University of Minnesota-Twin Cities Campus for his valuable input on statistical issues.

References

1. Hansch C (1976) J Med Chem 19:1–6
2. Basak SC, Gute BD, Grunwald GD (1999) A hierarchical approach to the development of QSAR models using topological, geometrical and quantum chemical parameters. In: Devillers J, Balaban AT (eds) Topological indices and related descriptors in QSAR and QSPR. Gordon and Breach, The Netherlands, pp 675–696
3. Basak SC, Grunwald GD, Niemi GJ (1997) Use of graph theoretical and geometrical molecular descriptors in structure–activity relationships. In: Balaban AT (ed) From chemical topology to three-dimensional geometry. Plenum, New York, pp 73–116
4. Basak SC (1987) Med Sci Res 15:605–609
5. Basak SC, Niemi GJ, Veith GD (1991) J Math Chem 7:243–272
6. Seydel JK, Schaper KJ, Wempe E, Cordes HP (1976) J Med Chem 19:483–492
7. Bagchi MC, Maiti BC (2003) J Mol Struct: THEOCHEM 623:31–37
8. Basak SC, Gute BD, Mills D (2002) SAR QSAR Environ Res 13:727-742
9. Basak SC, Mills D, Hawkins DM, El-Masri HA (2002) SAR QSAR Environ Res 13:649–665
10. Basak SC, Harriss DK, Magnuson VR (1988) POLLY, Version 2.3. Copyright of the University of Minnesota, USA
11. Filip PA, Balaban TS, Balaban AT (1987) J Math Chem 1:61–83
12. Basak SC, Balaban AT, Grunwald GD, Gute BD (2000) J Chem Inf Comput Sci 40:891-898
13. Hall Associates Consulting (2000) Molconn – Z, Version 3.50, Quincy, Mass.

14. Kier LB, Hall LH (1986) Molecular connectivity in structure–activity analysis. Research Studies Press, Letchworth, Hertfordshire, UK
15. Kier LB, Murray WJ, Randic M, Hall LH (1976) J Pharm Sci 65:1226–1230
16. Randic M (1975) J Am Chem Soc 97:6609–6615
17. Basak SC, Magnuson VR, Niemi GJ, Regal RR (1988) Discrete Appl Math 19:17–44
18. Balaban AT (1982) Chem Phys Lett 89:399–404
19. Balaban AT (1983) Pure Appl Chem 55:199–206
20. Balaban AT (1985) Math Chem (MATCH) 21:115–122
21. Raychaudhury C, Ray SK, Ghosh JJ, Roy AB, Basak SC (1984) J Comput Chem 5:581-588
22. Basak SC (1999) Information theoretic indices of neighborhood complexity and their applications. In: Devillers J, Balaban AT (eds) Topological indices and related descriptors in QSAR and QSPR. Gordon and Breach, The Netherlands, pp 563–593
23. Kier LB, Hall LH (1999) Molecular structure description: the electro topological state. Academic Press, San Diego, Calif.
24. Hall LH, Mohney B, Kier LB (1991) J Chem Inf Comput Sci 31:76–82
25. SAS Institute Inc (1988) In: SAS/STAT User Guide, Release 6.03 Edition. SAS Institute Inc, Cary, N.C.
26. Miller AJ (1990) Subset selection in regression. Chapman and Hall, New York
27. Rencher AC, Pun FC (1980) Technometrics 22:49–53
28. Hoerl AE, Kennard RW (1970) Technometrics 8:27–51
29. Massy WF (1965) J Am Statistical Assoc 60:234–246
30. Wold H (1975) Soft modeling by latent variables: The nonlinear iterative partial least squares approach. In: Gani J (ed) Perspectives in probability and statistics, papers in honor of Bartlett MS. Academic Press, London
31. Hoskuldsson AJ (1988) Chemometrics 2:211–228
32. Hoskuldsson AJ (1995) Chemometrics 9:91–123
33. Frank IE, Friedman JH (1993) Technometrics 35:109–135
34. Basak S, Mills D, Hawkins DM, El-Masri H (2002) Risk Anal (accepted)
35. Basak SC, Mills D, Mumtaz MM, Balasubramanian K (2003) Indian J Chem 42A:1385-1391
36. Hawkins DM, Basak SC, Mills D (2003) Environ Toxicol Pharmacol (accepted)
37. Hawkins DM, Yin X (2002) Comput Stat Data Anal 40:253–262
38. Hawkins DM, Basak SC, Mills D (2003) J Chem Inf Comput Sci 43:579–586