J Mol Model (2007) 13:111–120
DOI 10.1007/s00894-006-0133-z

ORIGINAL PAPER

Quantitative structure-activity relationship (QSAR) studies of quinolone antibacterials against *M. fortuitum* and *M. smegmatis* using theoretical molecular descriptors

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

Received: 15 February 2006 / Accepted: 28 June 2006 / Published online: 24 August 2006

Abstract The incidence of tuberculosis infections that are resistant to conventional drug therapy has risen steadily in the last decade. Several of the quinolone antibacterials have been examined as inhibitors of *M. tuberculosis* infection as well as other mycobacterial infections. However, not much has been done to examine specific structure–activity relationships of the quinolone antibacterials against mycobacteria. The present paper describes quantitative structure–activity relationship modeling for a series of antimycobacterial compounds. Most of the antimycobacterial compounds do not have sufficient 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 quantitative structure–activity relationship (QSAR) models utilizing theoretical molecular descriptors that can be calculated directly from molecular structures. Descriptors associated with chemical structures of N-1 and C-7 substituted quinolone derivatives as well as 8-substituted quinolone derivatives with good antimycobacterial activities against *M. fortuitum* and *M. smegmatis* have been evaluated. Ridge regression (RR), Principal component regression (PCR), and partial least squares (PLS) regression were used, comparatively, to develop predictive models for antibacterial activity, based on the activities of the above compounds. The independent variables include topostructural, topochemical and 3-D geometrical indices, which were used in a hierarchical fashion in the model-development process. The predictive ability of the models was assessed by the cross-validated R^2. Comparison of the relative effectiveness of the various classes of molecular descriptors in the regression models shows that the easily calculable topological indices explain most of the variance in the data.

Keywords Quantitative structure–activity relationships · Quinolone derivatives · Theoretical molecular descriptors · Ridge regression · Principal components regression · Partial least squares

M. C. Bagchi (✉)
Drug Design Development and Molecular Modelling Division,
Indian Institute of Chemical Biology,
4 Raja S.C. Mullick Road,
Calcutta 700032, Jadavpur, India
e-mail: mcbagchi@iicb.res.in

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

Introduction

The increase in the incidence of tuberculosis infections within the last decade can be attributed to a similar increase in the number of HIV infections. Numerous studies have shown that tuberculosis is a co-factor in the progression of HIV infections. Furthermore, tuberculosis has become resistant to conventional drug therapy, including isoniazide and rifampicin. Hence, there is a need to develop alternative chemotherapeutics for *Mycobacterium tuberculosis* infections. It has been observed that several of the quinolone antibacterials act as inhibitors of *M. tuberculosis* infections as well as other mycobacterial infections. Thus, a series of quinolone derivatives with substitutions at N-1 and C-7, as well as at the 8 positions, have been studied to examine the relationships between structural modifications and activities against *Mycobacterium fortuitum* and *Mycobacterium smegmatis* [1, 2]. The activities of these quinolone compounds against *M.*

fortuitum and *M. smegmatis* are considered primarily due to the fact that these two mycobacteria are used as a barometer of *M. tuberculosis* activity. However, not much has been done to examine specific structure–activity relationships of these quinolone antibacterials against *M. fortuitum* and *M. smegmatis*. The importance of these structure–activity relationship studies lies in development of predictive models and, since physicochemical data are not always available to develop predictive models, the only alternative is to utilize theoretical molecular descriptors which can be derived solely from the structure of the chemical compounds. The majority of theoretical descriptors are topological indices, which are numerical quantifiers of the molecular topology and encode information relating to size, shape, branching pattern, cyclicity and symmetry of the molecular graph. In tuberculostatic drug design and related QSAR research, the topological indices have been widely used [3–6].

The current study involves structure–activity relationships and the development of predictive models within the framework of three linear statistical methods. The topostructural (TS), topochemical (TC) and geometrical (3-D) indices associated with the chemical structures of N-1 and C-7 substituted quinolone derivatives, as well as 8 substituted quinolones, with good anti-mycobacterial activities against *M. fortuitum* and *M. smegmatis* have been evaluated. The activities of the quinolone compounds against these two organisms are used as a measure of anti-*M. tuberculosis* activity. Various linear regression methods, including ridge regression, principal component regression and partial least squares, based on topostructural, topochemical and geometrical indices have been utilized for the prediction of antibacterial activity against *M. fortuitum* and *M. smegmatis*.

Materials and methods

Structures of quinolone compounds and their biological activities

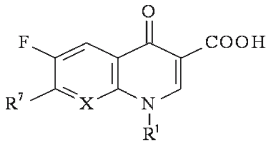
The activities of the quinolone compounds analyzed in this paper against *M. fortuitum* and *M. smegmatis* are shown in Table 1 [1, 2]. It has already been mentioned that the activity of the compounds against the rapidly proliferating organism *M. fortuitum* has been used as a measure of tuberculosis activity. This methodology is especially useful for the comparison of relative differences in activity between compounds. Biological activities against *M. smegmatis* are included for the purpose of comparison. Generally *M. fortuitum* is found to be more sensitive to the quinolone derivatives than *M. smegmatis*. The experimentally determined activity data are in the form of minimum inhibitory concentration (MIC in μg/ml) and these activities were considered by us for the construction of predictive QSAR models. Table 1 shows the structures of fluoroquinolones along with the side-chain substituents and activities against *M. fortuitum* and *M. smegmatis* considered in the present study.

Molecular descriptors

The molecular descriptors used in the present study are of 3 types: a) Topostructural (TS), b) Topochemical (TC) and c) Geometrical (3-D). All descriptors are based purely on molecular structure. Collectively, TS and TC descriptors are known as topological indices. The topostructural indices (TSIs) are calculated from the skeletal graph models of molecules, which do not distinguish among different types of atoms in a molecule or the various types of chemical bonds e.g., single bond, double bond, triple bond etc. Thus the TSIs quantify information regarding the connectivity, adjacency and distances between vertices, ignoring their distinct chemical nature. Topochemical indices (TCIs), on the other hand, are sensitive to both the pattern of connectedness of the vertices as well as their chemical and bonding characteristics. Therefore, the TCIs are more complex than the TSIs. The topological indices used in this study include a large set of connectivity indices [7–9], triplet indices [10, 11], electro-topological indices [12, 13], hydrogen bonding indices and information theoretic and neighborhood complexity indices [14, 15]. The 3-Dimensional or shape descriptors (3-D) encode information about the 3-dimensional aspects of molecular structures. The geometrical descriptors used in the present study include a set of Kappa shape indices [16, 17]. Using the hierarchical QSAR method, multiple models are developed, each time including an additional descriptor class that is more complex and computationally demanding. By comparing the statistical metrics of the hierarchically developed models, the relative contribution of each descriptor class can be examined.

The programs *Polly v 2.3* [18], *Triplet* [10, 11] and *Molconn-Zv3.5* [19] were used to calculate molecular descriptors in our present study. From *Polly* and associated software, a set of 102 topological descriptors is available including a large group of connectivity indices and path length descriptors [7–9, 20], Balaban's *J* indices [21–23] and information theoretic descriptors including neighborhood complexity indices [14, 15]. A set of 100 topological descriptors can be calculated from the *Triplet* program, where these descriptors 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

Table 1 Quinolone substrates and their activities against *M. fortuitum* and *M. smegmatis*



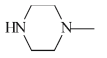
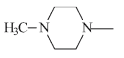
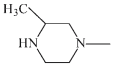
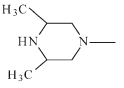
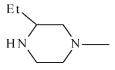
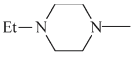
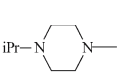
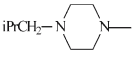
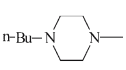
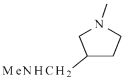
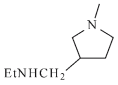
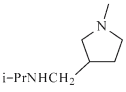
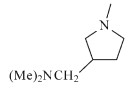
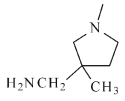
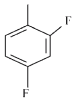
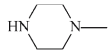
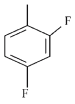
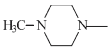
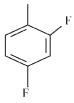
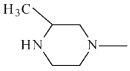
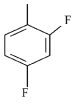
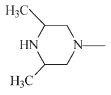
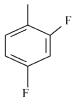
Comp No.	R^1	R^7	X	MIC Values(μg/mL) *M.fort*	MIC Values(μg/mL) *M.smeg*
1			CH	0.06	0.25
2			CH	0.06	0.25
3			CH	0.06	0.25
4			CH	0.06	0.13
5			CH	0.13	0.25
6			CH	0.06	0.13
7			CH	0.25	0.25
8			CH	1.0	0.25
9			CH	1.0	0.5
10			CH	0.03	0.25
11			CH	0.25	0.5
12			CH	0.5	0.5
13			CH	0.25	0.5
14			CH	0.13	0.5
15			CH	0.03	0.06
16			CH	0.25	0.5
17			CH	0.25	0.5
18			CH	0.13	0.5
19			CH	0.25	0.5
20			CH	0.25	1.0
21			CH	0.5	1.0
22			CH	0.5	1.0
23			CH	2.0	4.0
24			CH	1.0	2.0
25			CH	0.13	0.13
26			CH	0.13	0.25
27			CH	0.5	0.5
28			CH	0.5	2.0

Table 1 (continued)

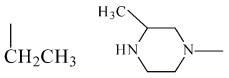
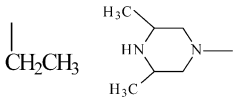
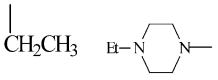
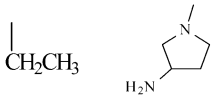
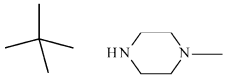
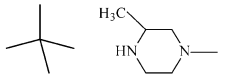
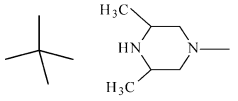
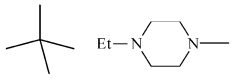
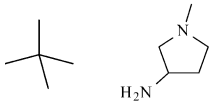
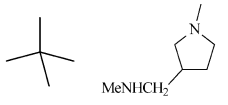
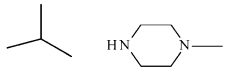
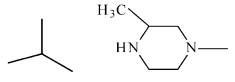
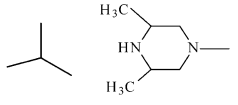
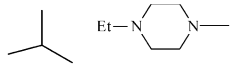
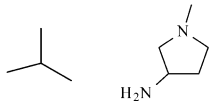
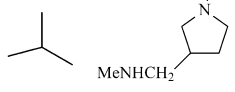
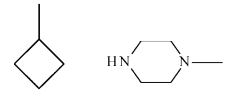
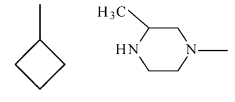
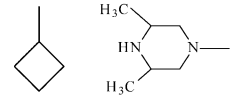
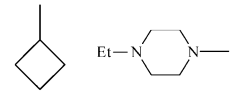
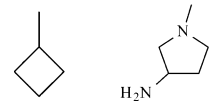
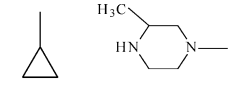
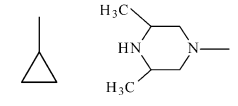
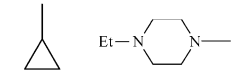
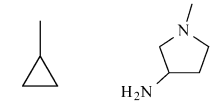
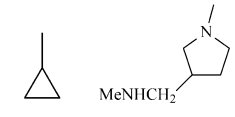
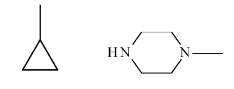
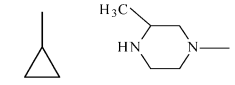
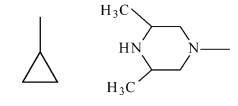
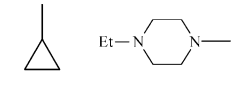
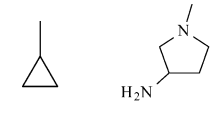
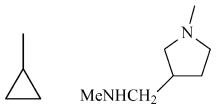
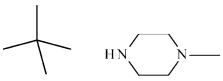
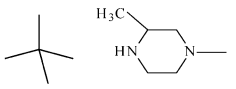
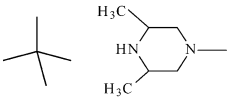
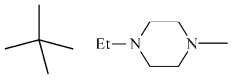
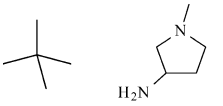
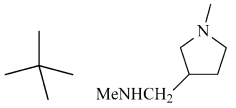
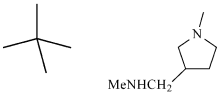
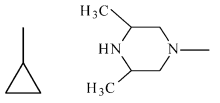
29			CH	1.0	2.0	30			CH	0.13	1.0
31			CH	0.25	0.5	32			CH	1.0	4.0
33			CH	2.0	8.0	34			CH	0.03	0.13
35			CH	0.03	0.06	36			CH	0.06	0.06
37			CH	0.13	0.13	38			CH	0.06	0.13
39			CH	0.13	0.25	40			CH	1.0	4.0
41			CH	0.5	2.0	42			CH	0.05	2.0
43			CH	1.0	4.0	44			CH	1.0	2.0
45			CH	2.0	8.0	46			CH	0.5	2.0
47			CH	0.25	1.0	48			CH	0.13	0.5
49			CH	0.5	2.0	50			CH	1.0	1.0
51			CBr	0.03	0.06	52			CBr	0.03	0.06
53			CBr	0.03	0.06	54			CBr	0.03	0.06
55			CBr	0.03	0.06	56			COMe	0.03	0.03
57			COMe	0.03	0.03	58			COMe	0.03	0.03
59			COMe	0.03	0.03	60			COMe	0.03	0.03

Table 1 (continued)

61			COMe	0.03	0.03	62			N	0.03	0.06
63			N	0.03	0.06	64			N	0.03	0.06
65			N	0.03	0.06	66			N	0.03	0.06
67			N	0.03	0.06	68			N	0.03	0.06
69[a]			CF	0.06	0.13						

[a]This structure contains an additional NH_2 group attached with Carbon at 5 position.

vertex invariants are then used in the following operations in order to obtain the *Triplet* descriptors:

(1) Summation, $\Sigma_i\, x_i$;
(2) Summation of squares, $\sum_i x_i^2$;
(3) Summation of square roots, $\sum_i x_i^{1/2}$;
(4) Sum of inverse square root of cross-product over edges ij, $\sum ij\left(x_i x_j\right)^{-1/2}$;
(5) Product, $N\left(\sum_i x_i\right)^{1/N}$.

A set of 167 descriptors including an extended set of connectivity indices, electrotopological indices and hydrogen bonding descriptors along with molecular shape descriptors were obtained from the *Molconn-Z* program. Table 2 shows the symbols, definitions and classifications of the theoretical molecular descriptors calculated for use in the present study.

In total, 369 molecular descriptors were calculated for each of the N-1, C-7 and 8 substituted quinolone compounds. However, it was found that 122 descriptors either had a constant value for all or nearly all of the compounds or were perfectly correlated with another descriptor, as identified by the CORR procedure of the SAS statistical package [24]. We have thus eliminated these 122 descriptors and reduced the total number of descriptors to 247, which were used for our predictive model development. Prior to model development, the 247 descriptors were transformed by the natural logarithm due to the fact that their scales differed by several orders of magnitude. The descriptors were partitioned into TS, TC, and 3D subsets and used in a hierarchical fashion in model development. The number of descriptors within each subset is 100, 140, and 7, respectively.

Statistical analysis

Prior to model development, the activity values were scaled by natural logarithm as their values differed by many orders of magnitude. Conventional ordinary least squares (OLS) regression cannot be used when the number of molecular descriptors exceeds the number of observations [25]. In this situation, three alternative linear regression methods may be considered, these are a) Ridge Regression (RR), b) Principal Component Regression (PCR) and c) Partial Least Squares (PLS). These three methods are also very useful even when the independent variables are highly correlated. In the ridge regression method, descriptors are transformed into principal components (PCs). All of the principal components are used in the regression, but they are first shrunk differentially according to their eigenvalues and a ridging constant. In the principal components regression, the descriptors are transformed into principal components after which a subset of the PCs is used in an ordinary least square regression. Partial least squares also uses a set of linear combinations of the descriptors but, in this approach, the dependent variable is also considered in this step. 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]. Statistical theory suggests that RR is the best of the three methods and this has been generally borne out in multiple

Table 2 Symbols, definitions and classification of structural molecular descriptors

Symbols	Definitions
Topostructural (TS)	
I_D^W	
$I^W{}_D$	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
H^{D}	Graph distance complexity
$\overline{IC}$	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
DN^2S_y	Triplet index from distance matrix, square of graph order, and distance sum; operation y=1–5
$DN^2{1}_y$	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
$DS1_y$	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
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
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
${}^h\chi^{\mathrm{b}}{}_C$	Bond cluster connectivity index of order h=3–6
${}^h\chi^{\mathrm{b}}{}_{Ch}$	Bond chain connectivity index of order h=3–6
${}^h\chi^{\mathrm{b}}{}_{PC}$	Bond path-cluster connectivity index of order h=4–6
${}^h\chi^{\mathrm{v}}$	Valence path connectivity index of order h=0–10
${}^h\chi^{\mathrm{v}}{}_C$	Valence cluster connectivity index of order h =3–6
${}^h\chi^{\mathrm{v}}{}_{Ch}$	Valence chain connectivity index of order h=3–10
${}^h\chi^{\mathrm{v}}{}_{PC}$	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

Table 2 (continued)

Symbols	Definitions
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; operation 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–Trinajstiæ information index
IdC	Bonchev–Trinajstiæ 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, $–NH_2$, –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)
$SHBint_y$	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, SssssSiH, SssssSi, SsPH2, SssPH, SssssP, 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
Geometrical (3D)/Shape	
kp_0	Kappa zero
kp_1–kp_3	Kappa simple indices
ka_1–ka_3	Kappa alpha indices

comparative studies [26–28]. As such, the RR models developed in the current study are analyzed in more detail than the PCR and PLS models The RR vector of regression coefficients, **b**, is given by

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

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. If k=0, RR reduces to conventional OLS regression. 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. This is due to the fact that the value of R^2 increases upon the addition of any descriptor, even when the descriptor is not relevant, and this results in overestimation of

model quality. Such an overestimation can be avoided by considering the cross-validated R^2, where the value of R^2 tends to decrease if some irrelevant descriptors are added, and this provides a reliable measure of model quality. Unlike R^2, the cross-validated R^2, denoted by R^2_{cv}, may be negative, signifying very poor model quality. The R^2_{cv} is calculated using the leave-one-out (LOO) approach, in which each compound is removed, in turn, from the data set and the regression is fitted based on the remaining n-1 compounds. For this reason, we have considered cross-validated regression models in our present study. R^2_{cv} is defined by

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

where SSTotal denotes 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 in this respect to using an external test set, it is a reliable measure of model predictability and this approach is preferred over the test-set approach when the sample size is small [29]. For comparative purposes, the conventional R^2 has also been reported for the RR models developed in this study. Comparison of these two metrics clearly shows the optimistic bias in conventional R^2. It is important to note again, however, that only R^2_{cv} is truly representative of the predictive quality of the models. The standard error of regression, calculated for the ridge regression models in this study, is also associated with the cross-validated, rather than the fitted, models.

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 important 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.

The TS, TC, and 3D-descriptor classes were used both alone and in a hierarchical fashion to develop multiple QSAR models using RR, PLS, and PCR, which could be compared in terms of their predictive ability.

Results and discussion

Preliminary analysis indicated compound #34 (Table 1) to be a statistical outlier based on model influence. As such, this compound was omitted. Results of the QSAR studies based on the remaining experimentally derived biological activity data of the N-1, C-7 as well as 8-substituted quinolone derivatives both for *M. fortuitum* and *M. smegmatis* are given in Table 3.

It is seen that the most predictive models are obtained using the simple topological indices with, for example, TS + TC resulting in a R^2_{cv} of 0.796 and 0.849 for *M. fortuitum* and *M. smegmatis*, respectively. The standard error of regression associated with these models is 0.53 and 0.41 $\log_e$ units, respectively. Considering the extremely large range of activity values within this data set, these standard error values are not out of line. From the information provided in Table 3, we can also conclude that the addition of the 3D descriptors does not improve the predictive power of the model. A comparison of the three statistical methods indicates that the PCR models are inferior to those developed using RR and PLS models. As stated previously, statistical theory suggests that RR is the best of the three methods, and previous studies have shown the superiority of the RR method [27, 28].

Table 3 QSAR analysis of quinolone antibacterials using RR, PCR & PLS models

Independent variables	RR			PCR	PLS
	R^2	R^2_{cv}	s.e. ($\log_e$ μg/ml)	R^2_{cv}	R^2_{cv}
M. fortuitum					
TS	0.903	0.760	0.61	0.566	0.742
TS+TC	0.900	0.796	0.53	0.489	0.792
TS+TC+3D	0.900	0.796	0.53	0.492	0.791
TS	0.903	0.760	0.61	0.566	0.742
TC	0.903	0.783	0.53	0.368	0.774
3D	0.526	0.446	1.00	0.433	0.468
M. smegmatis					
TS	0.971	0.799	0.60	0.595	0.775
TS+TC	0.967	0.849	0.41	0.482	0.854
TS+TC+3D	0.967	0.849	0.41	0.484	0.852
TS	0.971	0.799	0.60	0.595	0.775
TC	0.971	0.787	0.40	0.342	0.847
3D	0.612	0.444	1.12	0.444	0.434

Table 4 Important topological descriptors for the prediction of antimycobacterial activity based on *t* value, from the TS+TC ridge regression models

Descriptor label	Description	*t*
M. fortuitum		
AZN_4	Triplet index from adjacency matrix, atomic number, and graph order	−8.41
AZS_4	Triplet index from adjacency matrix, atomic number, and distance sum	−7.44
M_2	Sum of cross-product of degrees over all connected vertices	−7.37
P_2	Number of paths of length 2	−6.88
phia	Flexibility index	6.82
M_1	Sum of square of degrees over all vertices	−6.50
$^3\chi_{Ch}$	Chain connectivity index of order 3	−6.32
ANS_4	Triplet index from adjacency matrix, graph order, and distance sum	−5.93
ASV_3	Triplet index from adjacency matrix, distance sum, and vertex degree	−5.32
ANV_2	Triplet index from adjacency matrix, graph order, and vertex degree	−4.94
M. smegmatis		
SssO	Sum of the E-states for –O–	−5.47
NHBint5	Number of potential internal hydrogen bonds separated by 5 edges	5.14
$^3\chi_{Ch}$	Chain connectivity index of order 3	−5.08
$^6\chi_{Ch}$	Chain connectivity index of order 6	−4.89
SsF	Sum of the E-states for –F	4.81
phia	Sum of square of degrees over all vertices	4.77
P_5	Number of paths of length 5	−4.56
SssssC	Sum of the E-states carbon with four single bonds	4.26
AZN_4	Triplet from adjacency matrix, atomic number, and graph order	−4.03
SHBint2	Sum of E-state products for potential internal hydrogen bonds separated by 2 edges	3.98

It is instructive to look at the top ten molecular descriptors, based on t value, in the ridge regression models derived from TS + TC indices in Table 4. The models for the two organisms have some common descriptors; however, there are significant differences in the types of indices found to be important. This is due to the fact that for *M. fortuitum*, which is nearly as susceptible as *E. coli*, the corresponding MIC and IC_{50} values of quinolones are significantly lower than those found for *M. smegmatis*. In fact, the MIC values of quinolones against *M. fortuitum* are approximately two fold lower than those against *M. smegmatis* and the DNA gyrase from *M. fortuitum* displayed IC_{50} values two to eight fold lower than those from *M. smegmatis*. The correlation between activity against *M. fortuitum* and *M. smegmatis* and activity against *M. tuberculosis* is less good for the 8-substituted compounds than for ciprofloxacin and sparfloxacin.

Two statistical issues deserve further discussion. First, it should be clearly stated that, unlike ordinary least squares regression, the number of descriptors is not an issue with the regression methodologies used in this study. The number of descriptors included in the regression models are as follows: TS (100), TS+TC (240), TS+TC+3D (247), TC (140), 3D (7). These are appropriate methodologies when the number of descriptors exceeds the number of observations, and they are designed to use all available descriptors, as opposed to subset regression, in order to produce an unbiased model whose predictive ability is accurately reflected by the R^2_{cv}, regardless of the number of independent variables in the model. The distinction between these methods and OLS regression is important and cannot be overemphasized. Second, claims are often made that model validation through the use of an "external test set" is superior to the LOO approach. However, theoretical arguments and empirical studies [29] have shown that the LOO cross-validation approach is *preferred* to the use of an "external test set" unless the data set to be modeled is very large. The drawbacks of holding out an external test set include: 1) Structural features of the held out chemicals are not included in the modeling process, resulting in the 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 held-out set. The purpose of any validation procedure is to check the model fit independently of the model-fitting procedure; LOO cross-validation accomplishes this goal in a way that is far superior to other methods.

In contrast to models where the independent variables consist of experimentally determined physicochemical parameters and are essentially property-activity correlations, the models developed in this study use independent variables that are calculated molecular descriptors and, as such, are structure–activity correlations. The value of

QSAR modeling is evident when one considers the lack of experimental data available for modeling. The results of this study indicate that QSAR models using calculated molecular descriptors of quinolone antibacterials can be utilized in the design of alternative chemotherapeutics for *Mycobacterium tuberculosis* infection.

Acknowledgements Part of the research reported in this paper was supported 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. This paper represents contribution number 407 from the Center for Water and the Environment of the Natural Resources Research Institute. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Office of Scientific Research or the U.S. Government.

References

1. Reanau TE, Sanchiez JP, Gage JW, Dever JA, Shapiro MA, Gracheck SJ, Domagala JM (1996) J Med Chem 39:729–735
2. Reanau TE, Gage JW, Dever JA, Roland GE, Joannides ET, Shapiro MA, Sanchiez JP, Gracheck SJ, Domagala JM, Jacobs MR, Reynolds RC (1996) Antimicrob Agents Chemother 40:2363–2368
3. Bagchi MC, Maiti BC, Mills D, Basak SC (2004) J Mol Model 10:102–111
4. Bagchi MC, Maiti BC (2003) J Mol Struct: THEOCHEM 623:31–37
5. Bagchi MC, Maiti BC, Bose S (2004) J Mol Struct: THEOCHEM 679:179–186
6. Ghosh P, Thanadath M, Bagchi MC (2006) Molecular Diversity (in press). DOI 10.1007/s11030-006-9018-4
7. Kier LB, Hall LH (1986) Molecular connectivity in structure–activity analysis. Research Studies, Letchworth, Hertfordshire, UK
8. Kier LB, Murray WJ, Randic M, Hall LH (1975) J Pharm Sci 65:1226–1230
9. Randic M (1975) J Am Chem Soc 97:6609–6615
10. Filip PA, Balaban TS, Balaban AT (1987) J Math Chem 1:61–83
11. Basak SC, Balaban AT, Grunwald GD, Gute BD (2000) J Chem Inf Comput Sci 40:891–898
12. Kier LB, Hall LH (1990) Pharm Res 8:801–807
13. Kier LB, Hall LH (1999) Molecular structure description: The electrotopological state. Academic, San Diego, CA
14. Raychaudhury C, Ray SK, Ghosh JJ, Roy AB, Basak SC (1984) J Comput Chem 5:581–588
15. 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
16. Hall LH, Kier LB (1991) The molecular connectivity chi indexes and kappa shape indexes in structure–property relations. In: Boyd D, Lipkowitz K (eds) Reviews in computational chemistry. VCH, pp 367–422
17. Kier LB, Hall LH (1999) The kappa indices for modeling molecular shape and flexibility. In: Devillers J, Balaban AT (eds) Topological indices and related descriptors in QSAR and QSPR. Gordon and Breach Science, Amsterdam, pp 455–489
18. Basak SC, Harriss DK, Magnuson VR (1988) POLLY, Version 2.3. Copyright of the University of Minnesota, USA
19. Hall Associates Consulting (2000) Molconn-Z, Version 3.50. Quincy, Mass
20. Basak SC, Magnuson VR, Niemi GJ, Regal RR (1988) Discrete Appl Math 19:17–44
21. Balaban AT (1982) Chem Phys Lett 89:399–404
22. Balaban AT (1983) Pure Appl Chem 55:199–206
23. Balaban AT (1986) Math Chem (MATCH) 21:115–122
24. SAS Institute Inc (1988) SAS/STAT user guide, release 6.03 edn. SAS Institute, Cary, NC
25. Miller AJ (1990) Subset selection in regression. Chapman and Hall, New York, NY
26. Frank IE, Friedman JH (1993) Technometrics 35:109–135
27. Basak SC, Mills D, Hawkins DM, El-Masri H (2003) Risk Anal 23:1173–1184
28. Basak SC, Mills D, Mumtaz MM, Balasubramanian K (2003) Ind J Chem 42A:1385–1391
29. Hawkins DM, Basak SC, Mills D (2003) J Chem Inf Comput Sci 43:579–586