J Math Chem (2011) 49:185–200
DOI 10.1007/s10910-010-9734-y

ORIGINAL PAPER

Quantitative structure-activity relationship studies of boron-containing dipeptide proteasome inhibitors using calculated mathematical descriptors

Subhash C. Basak · D. Mills

Received: 6 May 2010 / Accepted: 13 September 2010 / Published online: 6 October 2010

Abstract Topological indices (TIs) and atom pairs (APs) were used to develop quantitative structure-activity relationship (QSAR) models of a set of 58 dipeptide boronic acids which are potent inhibitors of proteasome and have found applications in the treatment of various types of cancers. Of the three linear regression methods used for QSAR development, viz., principal components regression (PCR), partial least square (PLS), and ridge regression (RR), the last method gave the most satisfactory models whereas the remaining two methods yielded poor models. RR results obtained in this paper using TIs and APs are comparable to the CoMFA and CoMSIA results reported in the literature with the same set of compounds.

Keywords Quantitative structure-activity relationship (QSAR) · Topological indices · Atom pair · CoMFA and CoMSIA methods · Ridge regression

1 Introduction

In recent years, the proteasome has emerged as an important target for the design of chemotherapeutic agents. The 26S proteasome complex contains a 20S hollow catalytic unit capped by two regulatory 19S components [5]. The 20S subunit catalyzes the degradation of cellular proteins including those involved in cell signaling, control of ion channels, metabolic pathways, and responses to stress. Proteasome inhibitors have been found to induce apoptosis in rapidly dividing cells by interfering with the degradation of pro-growth cellular proteins [2]. Therefore, it is not surprising that

S. C. Basak (✉) · D. Mills
Natural Resources Research Institute, Center for Water and Environment, University of Minnesota Duluth, 5013 Miller Trunk Hwy, Duluth, MN 55811, USA
e-mail: sbasak@nrri.umn.edu

medicinal chemists have selected the 20S proteasome as a target for the design of novel chemotherapeutic agents [24,21,23].

It is interesting to note that, of all the proteasome inhibitors, dipeptide boronic acid derivatives are of special interest. One compound of this group (Velcade, known also as bortezomib) has been approved by the United States Food and Drug Administration (FDA) for the treatment of multiple myeloma patients for whom one prior cancer chemotherapy has failed and cases with relapsed mantle cell lymphoma. This compound is also being tested in Phase I, Phase II, and Phase III trials for the treatment of hematological malignancies and solid tumors [2,3].

Clinical data show that Velcade has various side effects including fatigue, nausea and sensory neuropathy [4]. Such deleterious effects might arise out of the fact that the 20S proteasome is involved in the breakdown of a multitude of cellular proteins which are critical in the maintenance of normal homeostatic mechanisms. One way to address the problem of side effects is to synthesize and test novel derivatives which might have more acceptable toxicity profiles. Zhu et al. carried out a quantitative structure-activity relationship (QSAR) analysis of boron-containing dipeptides using CoMFA and CoMSIA methods [40]. Such methods are, however, computationally demanding if one wishes to evaluate a large number of derivatives *in silico* before embarking on their actual synthesis in the laboratory.

Our research team at the Natural Resources Research Institute has been involved in the formulation of QSAR methods based on mathematical descriptors of ligands which can be easily calculated from their structure without the input of any other experimental data. QSARs based on such descriptors, viz., topological indices and substructures, have been useful in building successful predictive models for groups of chemicals working via receptor-based and enzymatic mechanisms [9,12,8]. Therefore, in this paper we used computed mathematical descriptors in developing QSARs for a set of 58 boron-containing dipeptides.

2 Methods and materials

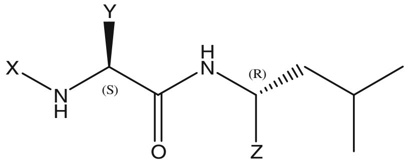
2.1 Data source

Dipeptide boronates with a common scaffold [40] were selected from the work of Adams et al. [1], since the biological activities of these compounds were measured under identical conditions. Chemicals with non-covalent bonds were excluded from the present study. The 58 chemicals in the data set, along with their biological activity values (p*Ki)*, are listed in Table 1.

2.2 Calculation of mathematical 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.

Table 1 Parent structure and substituents of 58 proteasome inhibitors, along with their experimentally determined activity values



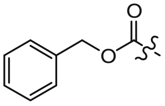
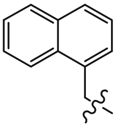
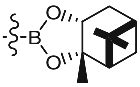
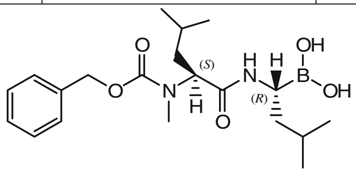
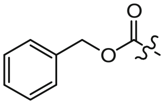
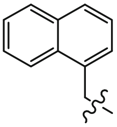
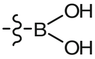
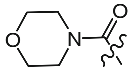
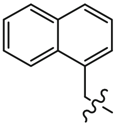
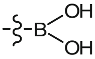
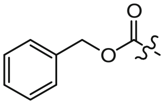
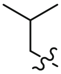
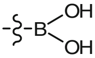
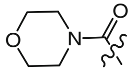
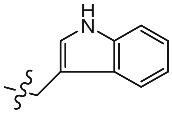
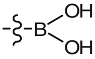
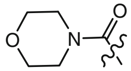
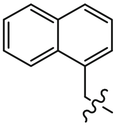
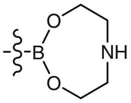
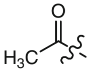
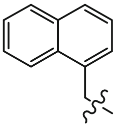
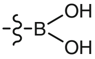
No.	X	Y	Z	*pKi*
1				10.00
2				6.00
3				10.08
4				9.74
5				8.52
6				8.52
7				9.82
8				9.77

Table 1 continued

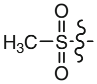
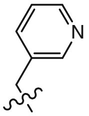
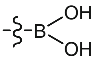
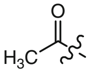
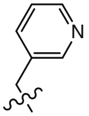
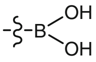
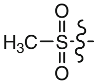
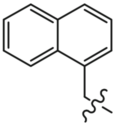
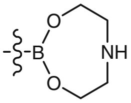
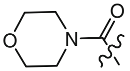
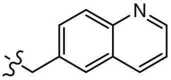
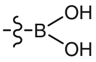
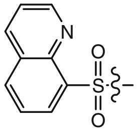
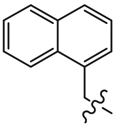
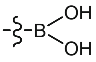
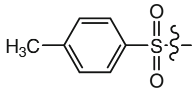
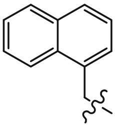
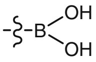
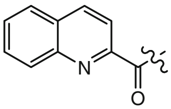
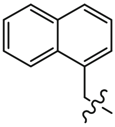
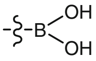
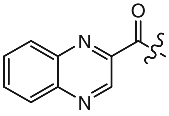
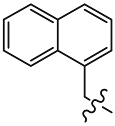
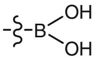
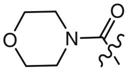
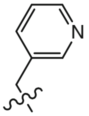
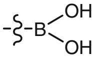
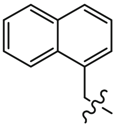
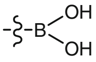
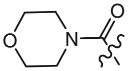
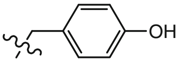
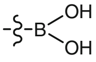
9				8.20
10				8.27
11				9.55
12				8.22
13				8.77
14				9.77
15				10.12
16				9.85
17				8.89
18	H			8.12
19				9.29

Table 1 continued

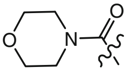
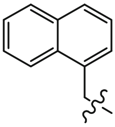
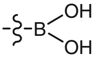
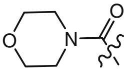
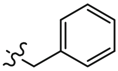
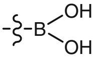
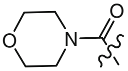
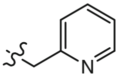
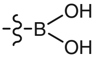
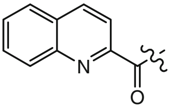
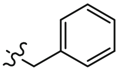
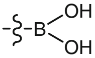
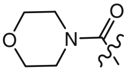
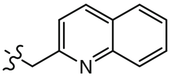
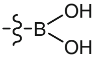
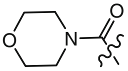
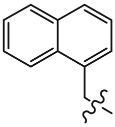
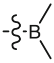
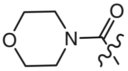
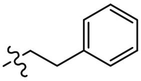
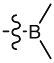
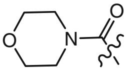
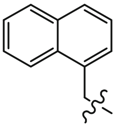
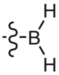
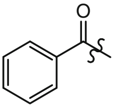
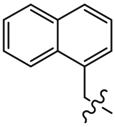
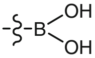
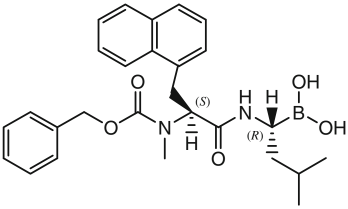
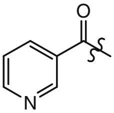
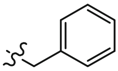
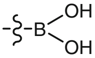
20				9.14
21				9.09
22				8.20
23				9.72
24				8.66
25				6.19
26				8.66
27				8.70
28				10.06
29				9.02
30				9.60

Table 1 continued

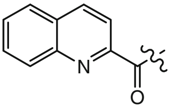
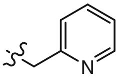
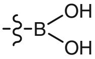
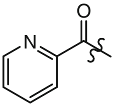
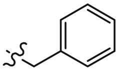
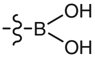
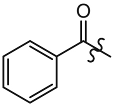
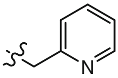
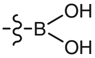
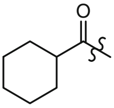
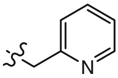
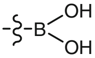
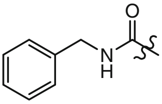
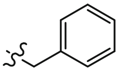
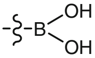
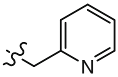
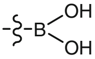
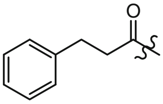
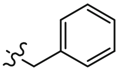
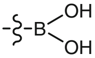
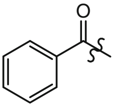
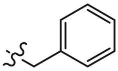
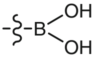
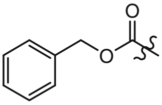
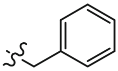
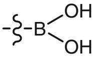
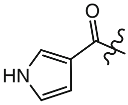
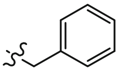
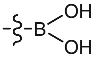
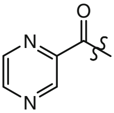
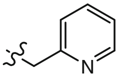
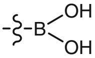
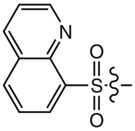
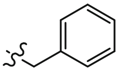
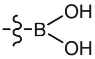
31				8.85
32				9.38
33				9.12
34				8.96
35				9.85
36	H			7.50
37				9.82
38				9.82
39				9.89
40				9.85
41				8.15
42				8.36

Table 1 continued

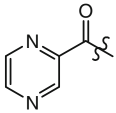
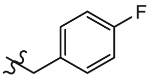
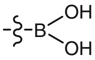
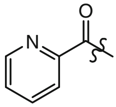
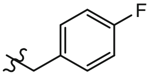
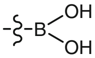
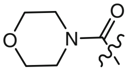
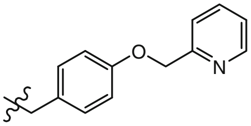
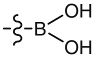
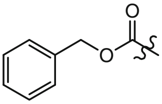
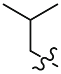
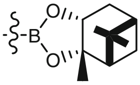
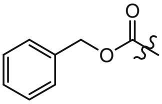
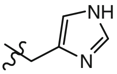
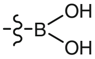
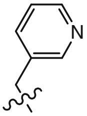
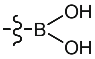
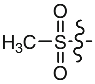
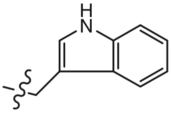
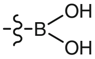
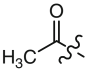
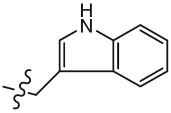
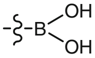
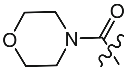
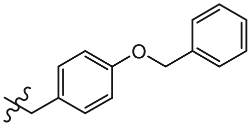
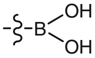
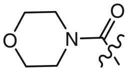
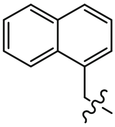
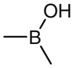
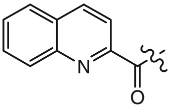
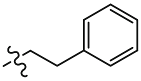
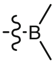
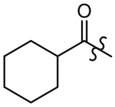
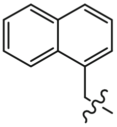
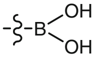
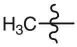
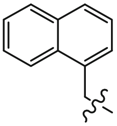
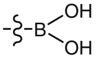
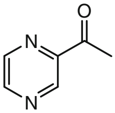
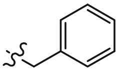
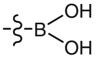
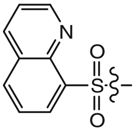
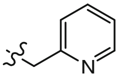
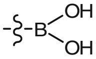
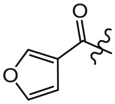
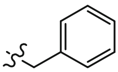
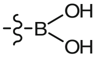
43				9.02
44				9.08
45				9.70
46				6.92
47				7.60
48	H			6.82
49				8.24
50				8.89
51				9.64
52				7.00
53				8.55

Table 1 continued

54				10.52
55				8.96
56				9.22
57				7.54
58				9.77

An atom pair represents any two atoms in the molecule and includes information about their path-wise interatomic separation as well as the presence or absence of π-orbitals. The method of Carhart et al. [22] 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:

$$< \text{atom descriptor } i > - < \text{separation} > - < \text{atom descriptor } j >$$

where <atom descriptor> contains information regarding atom type, number of non-hydrogen neighbors and the number of π 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 [18]. *APProbe* [7] was used to calculate the atom pairs for each molecule in the data set. In total, 805 APs were calculated for the boron-containing dipeptide data set.

In addition to the atom pairs, a set of 369 topological indices (TIs) was calculated using programs including *POLLY v2.3* [6], *Triplet* [25], and *Molconn-Z* [35]. The descriptors, based solely on chemical structure, include path length descriptors [32], path or cluster connectivity indices [32,36], neighborhood complexity indices [37], valence path connectivity indices [32], hydrogen bonding descriptors and electrotopological state indices [33]. Table 2 provides a list of the descriptors typically used in our studies, along with brief descriptions and hierarchical classification.

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
$\overline{I_D^W}$	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
$\overline{IC}$	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 $h = 0 - 10$
${}^h\chi_C$	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
$\mathrm{DN^2S}_y$	Triplet index from distance matrix, square of graph order, and distance sum; operation y=1 − 5
$\mathrm{DN^21}_y$	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 $y = 1 - 5$
$\mathrm{DN^2N}_y$	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 $y = 1 - 5$
$\mathrm{AN1}_y$	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 $y = 1 - 5$
kp_0	Kappa zero
$\mathrm{kp}_1 - kp_3$	Kappa simple indices

Table 2 continued

	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 *r*th ($r = 0 - 6$) order neighborhood of vertices in a hydrogen-filled graph
SIC_r	Structural information content for *r*th ($r = 0 - 6$) order neighborhood of vertices in a hydrogen-filled graph
CIC_r	Complementary information content for *r*th ($r = 0 - 6$) 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; 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-Trinajstić information index
IdC	Bonchev-Trinajstić information index
Wp	Wienerp

Table 2 continued

	Topochemical (TC)
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 ($y = 2 - 10$)
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, 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*

Prior to model development, any descriptor with a constant value for 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 [38] was retained. Subsequently, 248 TIs remained for use in the modeling study. Prior to modeling, the descriptors were standardized by autoscaling to zero mean and unit standard deviation.

3 Statistical analyses

Three regression methods that are appropriate when the number of descriptors exceeds the number of observations are ridge regression (RR) [28,29], principal component regression (PCR) [26], and partial least squares (PLS) regression [26,39]. 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 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 'shrinks' them differentially according to their eigenvalue [28]. 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 [31,30]. 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 [12,26,14,11,10,17]. The three methods were used comparatively 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:

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

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 [27] 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. [27] and Kraker et al. [34] 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 [34,19]. 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 statistical metric that has been examined in this review 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

Table 3 Regression results for models based on atom pairs (APs), topological indices (TIs), and the combination of APs and TIs

Descriptor.Set	q^2		
	RR	PCR	PLS
AP	0.682	−0.0673	0.492
TI	0.440	0.0080	−300.698
AP + TI	0.665	−0.0828	−24.245

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.

4 Results and discussion

The principal objective of this paper is to investigate how far easily calculated molecular descriptors such as topological indices and atom pairs are capable of developing QSAR for the class of boron-containing dipeptide proteasome inhibitors. Results shown in Table 3 show that these easily calculated molecular descriptors provide acceptable QSARs for this set of drug molecules. It is noteworthy that of the three statistical methods used for model building, viz., PCR, PLS, and RR, the best quality QSARs were obtained by the RR approach. This is in line with our earlier observations for various sets of molecules and QSARs of physicochemical, biochemical, toxicological, and pharmacokinetic properties [12,17,15,13,20,16].

It may be mentioned that our leave-one-out (LOO) q^2 values reported in Table 3 using either atom pairs alone (0.682) or a combination of TIs and APs (0.665) are comparable to those obtained by Zhu et al. [40] whose best LOO q^2 values were 0.676 and 0.630 using CoMFA and CoMSIA methods, respectively.

Although we obtained good quality QSARs for proteasome inhibitors using APs and TIs, it is important to understand the structural/ mechanistic basis of their activity. This can be done from an inspection of the descriptors that are influential in the QSAR models. Tables 4 and 5 list the descriptors with the highest $|t|$ values for the models derived by AP and AP plus TI, respectively.

As stated previously, each atom pair descriptor summarizes the connection path between two atoms. Thus, the AP labels in Tables 4 and 5 can be interpreted as follows, from left to right: Atomic symbol for 1st atom + pi electron count + "X" if non-hydrogen neighbors + # of non-hydrogen neighbors_interatomic separation_ Atomic symbol for 2nd atom + pi electron count + X" if non-hydrogen neighbors + # of non-hydrogen neighbors. The interatomic separation is the number of atoms traversed in the shortest bond-by-bond path containing both atoms. For example, N0X2_7_O0X1 represents a molecular fragment comprised of nitrogen and oxygen, with five intervening atoms; neither the nitrogen nor the oxygen is associated with any aromatic bonds; the nitrogen is bonded to two non-hydrogen neighbors, while the oxygen is bonded to one non-hydrogen neighbor.

Table 4 Atom pairs (APs) with high $|t|$ values, taken from the AP model

Atom pair	RRcoeff	s.e.	t	$\|t\|$
N0X2_7_O0X1	0.02567	0.00205	12.51	12.51
C0X1_3_C0X3	−0.01449	0.00129	−11.26	11.26
C0X1_4_C0X2	−0.01449	0.00129	−11.26	11.26
C0X1_5_C0X3	−0.03417	0.00307	−11.13	11.13
B0X3_2_C0X1	−0.05752	0.00524	−10.97	10.97
C0X1_4_N0X2	−0.05752	0.00524	−10.97	10.97
C0X1_9_O1X1	−0.05752	0.00524	−10.97	10.97
C0X1_6_C0X1	−0.03871	0.00354	−10.93	10.93
C0X1_6_C0X3	−0.05891	0.00551	−10.69	10.69
C1X3_8_O0X1	0.01652	0.00156	10.61	10.61

Table 5 Atom pairs (APs) and topological indices (TIs) with high $|t|$ values, taken from the AP+TI model

Atom pair	RRcoeff	s.e.	t	$\|t\|$
SsCH3	−0.00730	0.00064	−11.35	11.35
B0X3_2_C0X1	−0.05066	0.00447	−11.35	11.35
C0X1_4_N0X2	−0.05066	0.00447	−11.35	11.35
C0X1_9_O1X1	−0.05066	0.00447	−11.35	11.35
N0X2_7_O0X1	0.02073	0.00186	11.14	11.14
C0X1_3_C0X3	−0.01194	0.00109	−11.01	11.01
C0X1_4_C0X2	−0.01194	0.00109	−11.01	11.01
SsssB	−0.02778	0.00257	−10.82	10.82
C0X1_6_C0X3	−0.05086	0.00473	−10.76	10.76
C0X1_6_C0X1	−0.03480	0.00328	−10.61	10.61

The atom pair, N0X2_7_O0X1, with the highest $|t|$ value in Table 4 represents the boronic dipeptide substructure which is the essential backbone of the compounds, and the same is true of B0X3_2_C0X1, another influential substructure in Table 4. A number of APs have the methyl substituent in one or both ends showing the importance of hydrophobic methyl groups in the bioactivity of the boron containing dipeptides. The methyl groups may also indicate the importance of steric factors in the bioactivity of boronic acids. The CoMFA analysis of Zhu et al. [40] using the same set of compounds found steric factor to have considerable importance.

A perusal of results in Table 5 shows that the majority of the descriptors in the AP+TI model with high $|t|$ values are atom pairs rather than topological indices. The two TIs included in Table 5 are electrotopological state indices for the methyl group and boron, respectively. Here SsCH3, the electrotopological state index for the methyl group, emerged as an influential descriptor. The electrotopological index, SsssB, indicates the critical importance of boron. The boron-carbon substructure, B0X3_2_C0X1, appears to be important here as in the case of the AP-only QSAR analyses. Multiple substructures in Table 5 indicate the critical importance of methyl groups. Overall, the influential independent variables of Table 5 indicate that steric/ hydrophobic factors represented by methyl group containing substructures and specific electronic factors represented by boron are important in the pharmacological action of the dipeptide boronic acids. This is in line with the findings of Zhu et al. [40] using CoMFA and CoMSIA that steric and electrostatic factors play dominant roles in the mode of action of the boronic acid proteasome inhibitors analyzed in this paper.

In conclusion, easily calculated descriptors such as atom pairs and topological indices are capable of producing high-quality QSARs for boron-containing dipeptide proteasome inhibitors. Of the three statistical methods used, RR gave the best models. An inspection of the descriptors influential in the QSARs indicate that steric and electronic factors play dominant roles in determining bioactivity of the boron-containing dipeptide proteasome inhibitors. In view of the fact that the QSARs reported in this paper are based on purely mathematical structural descriptors which can be calculated very quickly, the models can help in practical drug design via rapid screening of large real or virtual libraries of boron-containing dipeptide derivatives.

Acknowledgments This is contribution number 506 from the Center for Water and the Environment of the Natural Resources Research Institute.

References

1. J. Adams, Y.T. Ma, R. Stein, M. Baevsky, L. Grenier, L. Plamonda, PCT Int. Appl., WO 1996013266. (1995)
2. J. Adams, V.J. Palombella, E.A. Sausville, J. Johnson, A. Destree, D.D. Lazarus, J. Maas, C.S. Pien, S. Prakash, P.J. Elliott, Cancer Res. **59**, 2615 (1999)
3. J. Adams, Trends Mol. Med. **8**(Suppl), S49–S54 (2002)
4. J. Adams, Drug Discov. Today **8**, 307 (2003)
5. J.B. Almond, G.M. Cohen, Leukemia **16**, 433 (2002)
6. S.C. Basak, D.K. Harriss, V.R. Magnuson, POLLY v. 2.3, Copyright of the University of Minnesota, 1988
7. S.C. Basak, G.D. Grunwald, APProbe. Copyright of the University of Minnesota, 1993
8. S.C. Basak, B.D. Gute, in *Proceedings of the International Congress on Hazardous Waste: Impact on Human and Ecological Health*, ed. by B.L. Johnson, C. Xintaras, J. S. Andrews (Princeton Scientific Publishing Co, Inc., 1997), pp. 492–504
9. S.C. Basak, B.D. Gute, S. Ghatak, J. Chem. Inf. Comput. Sci. **39**, 255 (1999)
10. S.C. Basak, D. Mills, D.M. Hawkins, H.A. El-Masri, SAR QSAR Environ. Res. **13**, 649 (2002)
11. S.C. Basak, D. Mills, D.M. Hawkins, H. El-Masri, Risk Analysis **23**, 1173 (2003)
12. S.C. Basak, D. Mills, M.M. Mumtaz, K. Balasubramanian, Indian J. Chem. **42A**, 1385 (2003)
13. S.C. Basak, D. Mills, B.D. Gute, D.M. Hawkins, in *Quantitative Structure-Activity Relationship* (*QSAR*) *Models of Mutagens and Carcinogens*, ed. by R. Benigni (CRC Press, Boca Raton, FL, 2003), pp. 207–234
14. S.C. Basak, D. Mills, H.A. El-Masri, M.M. Mumtaz, D.M. Hawkins, Environ. Toxicol. Pharmacol. **16**, 45 (2004)
15. S.C. Basak, D. Mills, ARKIVOC 2005, 308 (2005)
16. S.C. Basak, D. Mills, ARKIVOC 2005, 60 (2005)
17. S.C. Basak, D. Mills, B.D. Gute, SAR QSAR Environ. Res. **17**, 515 (2006)
18. S.C. Basak, B.D. Gute, D. Mills, ARKIVOC 2006, 157 (2006)
19. S.C. Basak, R. Natarajan, D. Mills, D.M. Hawkins, J.J. Kraker, J. Chem. Inf. Model. **46**, 65 (2006)
20. S.C. Basak, D. Mills, B.D. Gute, R. Natarajan, in *Topics in Heterocyclic Chemistry*, ed. by S.P Gupta Vol. 5: QSAR and Molecular Modeling Studies of Heterocyclic Drugs (Springer, Berlin-Heidelberg-New York, 2006), pp. 39–80

21. E.J. Corey, W.-D.Z. Li, T. Nagamitsu, G. Fenteany, Tetrahedron **55**, 3305 (1999)
22. R.E. Carhart, D.H. Smith, R. Venkataraghavan, J. Chem. Inf. Comput. Sci. **25**, 64 (1985)
23. J.G. Delcros, M.D. Floc'h, C. Prigent, Y. Arlot-Bonnemains, Curr. Med. Chem. **10**, 479 (2003)
24. G. Fenteany, R.F. Standaert, W.S. Lane, S. Choi, E.J. Corey, S.L. Schreiber, Science **268**, 726 (1995)
25. P.A. Filip, T.S. Balaban, A.T. Balaban, J. Math. Chem. **1**, 61 (1987)
26. I.E. Frank, J.H. Friedman, Technometrics **35**, 109 (1993)
27. D.M. Hawkins, S.C. Basak, D. Mills, J. Chem. Inf. Comput. Sci. **43**, 579 (2003)
28. A.E. Hoerl, R.W. Kennard, Technometrics **12**, 55 (1970)
29. A.E. Hoerl, R.W. Kennard, Technometrics **12**, 69 (2005)
30. A. Hoskuldsson, J. Chemometrics **2**, 211 (1988)
31. A. Hoskuldsson, J. Chemometrics **9**, 91 (1995)
32. L.B. Kier, L.H. Hall, *Molecular Connectivity in Structure-Activity Analysis* (Research Studies Press, Letchworth, Hertfordshire, U.K., 1986)
33. L.B. Kier, L.H. Hall, *Molecular Structure Description: The Electrotopological State* (Academic Press, San Diego, CA, 1999)
34. J.J. Kraker, D.M. Hawkins, S.C. Basak, R. Natarajan, D. Mills, Chemometr. Intell. Lab. Syst. **87**, 33 (2007)
35. Molconn-Z Version 3.5, Hall Associates Consulting, Quincy, MA., 2000
36. M. Randic, J. Am. Chem. Soc. **97**, 6609 (1975)
37. A.B. Roy, S.C. Basak, D.K. Harriss, V.R. Magnuson, in *Mathematical Modelling Science Technology*, ed. by X.J.R. Avula, R.E. Kalman, A.I. Liapis, E.Y. Rodin (Pergamon Press, New York, 1983), pp. 745–750
38. SAS Institute, Inc., *In SAS/STAT User Guide, Release 6.03 Edition* (SAS Institute Inc., Cary, NC, 1988)
39. S. Wold, Technometrics **35**, 136 (1993)
40. Y.-Q. Zhu, M. Lei, A.-J. Lu, X. Zhao, X.-J. Yin, Q.-Z. Gao, Eur. J. Med. Chem. **44**, 1486 (2009)