Classification of Inhibitors of Protein Tyrosine Phosphatase 1B Using Molecular Structure Based Descriptors

S. J. Patankar and P. C. Jurs*

Department of Chemistry, 152 Davey Laboratory, Penn State University, University Park, Pennsylvania 16802

Received June 24, 2002

Loss of Protein Tyrosine Phosphatase 1B (PTP 1B) activity is known to enhance insulin sensitivity and resistance to weight gain. So potent and orally active PTP1B inhibitors could be potential pharmacological agents for the treatment of Type 2 diabetes and obesity. Classification models of PTP1B inhibitors are developed using a data set containing 128 compounds. Their inhibitory concentrations ranged from −1.59 to 1.68 log units. Initially a two-class (active, inactive) problem is tackled using a number of different methods. The data set was divided into active and inactive classes on the basis of inhibitory activity of the compounds. Molecular structure-based descriptors were calculated and used in the model development. Descriptors encoding the flexibility of the molecules were investigated. Classification models were generated using k-nearest neighbors (k-NN), linear discriminant analysis (LDA), and radial basis function neural network (RBFNN). All models are tested using an external prediction set, compounds not used anywhere during the model development procedure. A five-descriptor model is developed that produces a classification rate of 85.7% for an external prediction set. Then a three-class (active, moderately active, inactive) problem was explored. This time the data set was divided into highly active, moderate, and inactive classes on the basis of inhibitory activity of the compounds. The best classification rate achieved for an external prediction set was 85%. The classification rates achieved indicate that these models could serve as a screening mechanism, to identify potentially useful PTP 1B inhibitors. In addition multiple linear regression and computational neural network models are also developed for prediction of log IC_{50} values. All QSAR models are tested using the same external prediction set.

INTRODUCTION

With the increase of obesity in the general population, especially young adults, the prevalence of diabetes is also on the rise.[1] This has led to a significant amount of interest in ways to control this disease. Diet and exercise can help in reducing the blood sugar; however, in many cases the sugar never comes down to the normal levels. Insulin regulates the body's ability to transport and use glucose.[1] Insulin resistance along with lack of insulin production, are common causes of Type 2 diabetes.[2] Diabetes can lead to a number of debilitating complications.[3] Diet, exercise, and the use of hypoglycemic agent is recommended for reducing the blood sugar. It is noted that patients develop resistance to currently available hypoglycemic agents after a number of years of treatment.[4−6]

Recently there is a considerable interest in the intracellular phosphatase, protein tyrosine phosphatase 1B (PTP 1B),[7,8] as it plays a major role in the dephosphorilation of insulin receptors.[9,10] PTP 1B acts as a negative regulator of insulin receptors.[11] Therefore loss of PTP 1B activity enhances insulin sensitivity and resistance to weight gain. So the compounds that inhibit PTP 1B activity can be useful in the treatment of Type 2 diabetes and obesity. Several references describing peptide and small molecule inhibitors of PTP 1B are available.[12−14]

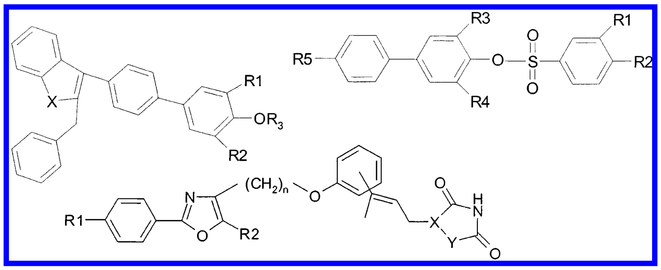
In this paper we have used computer modeling to classify a structurally varied set of PTP 1B inhibitors with known antihyperglycemic properties. This data set consists of 10 scaffolds of various azolidinediones and benzofuran benzothiophene biphenyls. Classification of the actives/inactives as well as actives/moderates/inactives was done on the basis of molecular structure alone. The predictive ability of all models developed is examined using an external prediction set. This modeling technique can be used even when complete knowledge of the binding site is unavailable. Models developed could be used to screen large libraries of compounds to identify those likely to display activity as PTP 1B inhibitors.

Along with classification we have used quantitative structure activity relationship (QSAR) to predict the inhibitory concentrations of this varied data set. Prediction of an inhibitory concentration was done on the basis of molecular structure alone. This type of methodology has significant benefits such as cost saving, safety, lack of consumption of test sample, and shortened time, etc.

EXPERIMENTAL SECTION

This study was performed on a combined data set taken from two papers by Malamas et al.[15,16] To maximize the likelihood of finding a relationship between chemical structure and PTP 1B inhibitory activity, only the data collected under the same experimental conditions was used. The combined data set included 128 compounds, out of which 12 were beyond the software limits. The 128 compounds were variations of 10 different moieties, three of which are shown below.

10.1021/ci020045e CCC: $25.00
Published on Web 04/05/2003



The compounds were evaluated by Malamas et al.[15,16] for their in vitro activity against recombinant human[17] PTP1B with phosphotyrosil dodecapeptide TRDI(P)-YETD(P)Y(P)-YRK (corresponds to the 1142−1153 insulin receptor kinase regulatory domain, phosphorylated on the 1146, 1150, 1151 tyrosine residues; IR − triphosphopeptide) as the source of the substrate.[18] Enzyme reaction progression was monitored via the release of inorganic phosphate as detected by the malachite − green ammonium molybdate method.[19] The in vitro activity was expressed as the micromolar concentration of the test compound which inhibited enzyme activity by 50% (IC_{50}).

The molecular weight for the data set compounds varied from 131 to 814 amu with a mean of 552 amu. The log IC_{50} values ranged from −1.59 to 1.68 log(μM) units. All compounds contained oxygen, 58 compounds contained sulfur, and 36 compounds contained nitrogen. Overall 59 compounds contained halogens, out of which 36 contained bromine, 21 contained fluorine, 3 contained iodine, and 2 contained chlorine. Additionally, all compounds were aromatics with anywhere from 4 to 7 rings. Structural information for all the compounds is given in Table 1.

Since this study reports results for classification as well as QSAR work from hereon all the work is discussed under two separate headings. The common portions will be discussed under classification study and will be referenced under the QSAR study.

Classification. To generate two-class models the data set was divided into active and inactive compounds on the basis of their activity. As there were no reported guidelines on this data set to use for the required subdivisions, compounds with log $IC_{50} \leq -0.7$ were considered active and compounds with log $IC_{50} > -0.7$ were considered inactive. For the three-class problem log $IC_{50} \leq -1.2$ were active, $-1.2 >$ log $IC_{50} \leq -0.7$ were moderately active, and log $IC_{50} > -0.7$ were considered inactive. As the same external prediction set is used throughout the study the cutoffs were chosen such that there is some representation of actives and moderates as well as inactives in all subsets. As the data set is very small different cutoff points were not investigated. At this point a set of 14 compounds was randomly selected as a prediction set (PSET), and these compounds were not used anywhere in the model development but for validation. The remaining 114 compounds comprised the training (TSET). Table 2 shows the distribution of compounds into classes for the data set.

Methods used to develop classification models can be broken down to four basic steps: structure entry, descriptor generation, objective feature selection, and model development and validation. Several quantitative structure activity relationship studies have reported similar type of methodology previously.[20−23]

Structure Entry. The compounds were sketched as 2-D representations using HyperChem,[24] and optimized 3-D conformations were generated. Compound structures were stored as connection tables, which contain information about atom types, bond angles, and bond hybridizations. They were used to generate topological descriptors. At a later stage these structures were further refined to their lowest energy states using MOPAC,[25] a semiempirical molecular modeling routine. A PM3 Hamiltonian was selected for geometry optimization.[26] These optimum 3-D conformations were used for generation of descriptors dependent on geometry.

Descriptor Generation. To relate molecular structure to PTP 1B inhibitory activity, descriptors that accurately encode the structural features responsible for the observed activity are necessary. The ADAPT (**A**utomated **D**ata **A**nalysis and **P**attern Recognition **T**oolkit) software package[27,28] was used to calculate approximately 100 descriptors for each compound. As some of the compounds were beyond software limits, descriptor generation routines were modified to accommodate the larger molecules. This led to almost 50% reduction in the number of descriptors generated. The calculated descriptors encode the geometric, topological, electronic features and flexibility[29] of the compounds.

Topological descriptors[30−33] are calculated on the basis of a two-dimensional sketch of the compound. A significant advantage of these type of descriptors is that geometry optimization of the structure is not needed. Topological descriptors calculated included connectivity indices, molecular distance edge descriptors, kappa indices, and flexibility indices. Connectivity indices and molecular distance edge descriptors encoded information about molecular size and branching. Kappa indices provided information about molecular shape using the two-dimensional structure. The new flexibility indices routine provided collective information about the flexibility of the molecule and mass equivalence of the rotatable as well as rigid atoms.

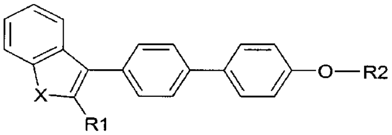
Geometric descriptors[34,35] such as volume, solvent accessible surface areas, and gravitational indices provide information about size and ability of the compounds to interact with solvents on a 3-D basis. However these could not be calculated as some of the compounds in the data set were beyond the software limit.

Electronic descriptors[36] encode the electronic environment of the compounds, and they are frequently computationally demanding. These descriptors were calculated after geometry optimization of all the compounds. Our own as well as MOPAC routines were developed to accommodate the larger molecules in the data set. Calculated descriptors included dipole moment, polarizability, electronegativity, energy of the highest occupied molecular orbital, and the energy of the lowest unoccupied molecular orbital. However due to the software limit CPSA[36] descriptors were not calculated.

Objective Feature Selection. The descriptors were subdivided into two pools, one containing only topological descriptors and the other containing all descriptors, topological as well as electronic descriptors. The next step involved objective feature selection. The process of feature selection entails pruning the descriptor pool through objective and subjective means. Objective feature selection eliminates descriptors based solely on their values. In this process the dependent variable information (class label) is not utilized. To avoid chance correlation the descriptor pool was reduced

Table 1. Structures, Log IC$_{50}$, and True and Predicted Classes for PTP 1B Inhibitors

Benzofuran and Benzothiophene Biphenyls



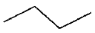
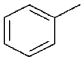
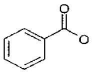
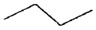
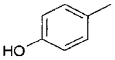
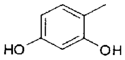
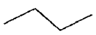
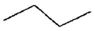
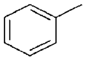
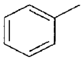
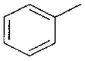
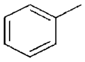
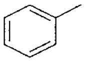
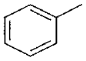
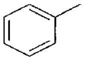
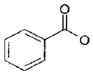
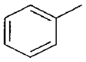
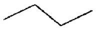
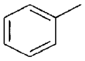
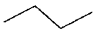
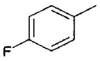
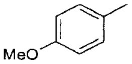
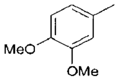
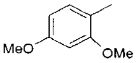
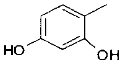
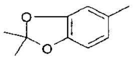
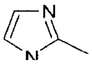
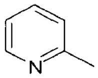
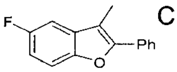
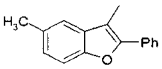
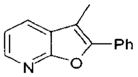
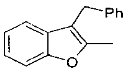
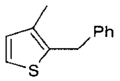
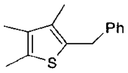
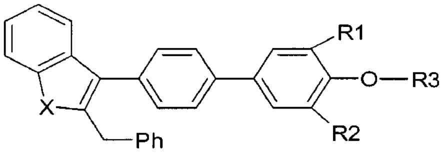
Comp. No.	R1	R2	X	Log IC$_{50}$	Two Class Problem		Three Cls
					TRUE Class	Pred.[a] Class	Pred.[b] Class
1		H	O	-0.1308	1	1	1
2		H	O	-0.0362	1	1	1
3		H	O	-0.1308	1	1	1
4[c]		H	S	-0.1549	1	1	1
5		H	S	0.0334	1	1	1
6		H	S	-0.2366	1	1	1
7		CH_2CO_2H	O	0.3404	1	1	1
8[d]		$CH(CH_2Ph)CO_2H$	O	-0.3565	1	2	1
9		$CH(CH_2Ph)CO_2H$	O	-0.5686	1	1	1
10		$CH(CH_2Ph)CO_2H$(R)	O	-0.4559	1	1	2
11		$CH(CH_2Ph)CO_2H$(S)	O	-0.4949	1	1	1
12		$CH(CH_2Ph)CO_2H$(S)	O	-0.6576	1	2	1
13		$CH(CH_2Ph)CO_2H$	O	-0.5376	1	1	1
14		$CH(CH_2Ph)CO_2H$(R)	O	-0.3979	1	1	1
15		$CH(CH_2Ph)CO_2H$(R)	O	0.1206	1	1	1
16		$CH(CH_2Ph)CO_2H$(R)	O	-0.1675	1	1	2
17	CH(OH)Ph	$CH(CH_2Ph)CO_2H$(R)	O	-1.0000	2	1	1
18		$CH_2Ph\text{-}_4\text{-}CO_2H$	O	-0.4437	2	2	1
19[d]		$CH(CH_2Ph)CO_2H$(R)	S	-0.7696	2	2	2
20		$CH(CH_2Ph)CO_2H$(R)	S	-1.0223	2	1	2
21		$CH(CH_2Ph)CO_2H$(R)	S	-0.9586	2	2	2
22[c]		$CH(CH_2Ph)CO_2H$(R)	S	-0.9208	2	2	3

Table 1. (Continued)

Comp. No.	R1	R2	X	Log IC_{50}	Two Class Problem		Three Cls
					TRUE Class	Pred.[a] Class	Pred.[b] Class
23		$CH(CH_2Ph)CO_2H(R)$	S	-1.1135	2	2	2
24		$CH(CH_2Ph)CO_2H(R)$	S	-0.9208	2	2	2
25		$CH(CH_2Ph)CO_2H(R)$	S	-1.0706	2	2	2
26		$CH(CH_2Ph)CO_2H(R)$	S	-0.9208	2	2	2
27		$CH(CH_2Ph)CO_2H(R)$	S	-1.1135	2	2	2
28[d]		$CH(CH_2Ph)CO_2H(R)$	S	0.0645	2	2	2
29		$CH(CH_2Ph)CO_2H(R)$	S	0.1903	1	1	1
30[d]		$CH(CH_2Ph)CO_2H(S)$	C	-0.8861	2	1	2
31		$CH(CH_2Ph)CO_2H(R)$	C	-0.3872	1	1	1
32[c]		$CH(CH_2Ph)CO_2H(S)$	C	-0.2291	1	1	1
33		$CH(CH_2Ph)CO_2H(R)$	C	-0.4559	1	1	1
34		$CH(CH_2Ph)CO_2H(R)$	C	-0.0132	1	1	1
35		$CH(CH_2Ph)CO_2H(R)$	C	-0.2924	1	1	1

2-Benzyl Benzofuran and Benzothiophene Biphenyls



Comp. No.	R1	R2	R3	X	Log IC_{50}	Two Class Problem		Three Cls
						TRUE Class	Pred.[a] Class	Pred.[b] Class
36	Br	H	H	S	0.0294	1	1	1
37	Br	Br	H	S	-0.3468	1	1	1
38	I	I	H	S	-0.2840	1	1	1
39	Br	H	$CH(CH_2Ph)CO_2H(R)$	S	-1.2366	2	2	3
40	Br	Br	$CH(CH_2Ph)CO_2H(R)$	S	-1.6021	2	2	3
41	4-OCH_3-Ph		$CH(CH_2Ph)CO_2H(R)$	S	-1.2757	2	2	3
42[c]	4-Cl-Ph	H	$CH(CH_2Ph)CO_2H(R)$	S	0.4771	2	2	3
43	Br	Br	$CH(CH_2CH_2Ph)CO_2H(S)$	S	-0.5376	1	2	3

Table 1. (Continued)

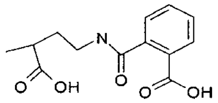
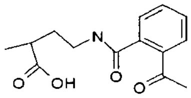
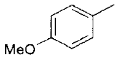
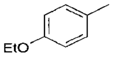
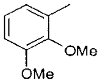
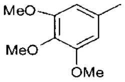
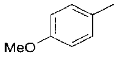
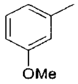
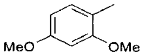
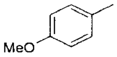
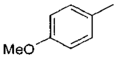
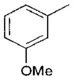
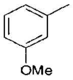
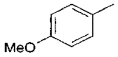
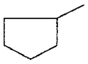
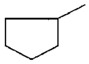
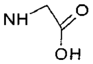
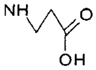
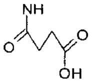
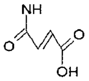
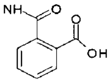
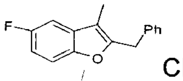
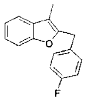
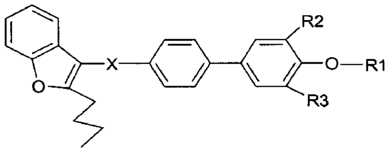
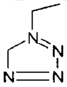
Comp.						Two Class Problem		Three Cls
No.	R1	R2	R3	X	Log IC_{50}	TRUE	Pred.[a]	Pred.[b]
						Class	Class	Class
44	Br	Br		S	0.4771	2	2	3
45	Br	Br		S	-1.2676	2	2	3
46[d]	Br	H	CH_2CO_2H	S	-0.4437	1	2	1
47	Br	Br	CH_2CO_2H	S	-1.0000	2	2	2
48	Ph	H	CH_2CO_2H	S	-1.0000	2	2	2
49		H	CH_2CO_2H	S	-1.0969	2	2	2
50[d]		H	CH_2CO_2H	S	-1.2840	2	2	2
51		H	CH_2CO_2H	S	-1.1487	2	2	2
52		H	CH_2CO_2H	S	-1.0000	2	2	2
53		Br	CH_2CO_2H	S	-1.5376	2	2	3
54		Br	CH_2CO_2H	S	-1.5528	2	2	3
55[c]		Br	CH_2CO_2H	S	-1.3279	2	2	3
56			CH_2CO_2H	S	-1.6021	2	2	3
57			CH_2CO_2H	S	-1.6021	2	2	3
58	Br	H		S	-0.7696	2	2	1
59	Br	H	$CH(CH_2Ph)CO_2H(S)$	O	-1.2518	2	2	3
60	Br	Br	$CH(CH_2Ph)CO_2H(S)$	O	-1.4202	2	2	3
61		H	$CH(CH_2Ph)CO_2H(S)$	O	0.4771	2	2	3
62	NO_2	H	$CH(CH_2Ph)CO_2H(R)$	O	-0.6383	1	1	1
63	Br	Br	$CH(CH_2CH_2Ph)CO_2H(S)$	O	-0.8861	2	2	3
64	Br	Br	$CH[CH_2CH(CH_3)_2]CO_2H(R)$	O	-1.2676	2	2	3
65	Br	Br	$CH[(CH_2)_3CH_3]CO_2H$	O	-1.2840	2	2	3
66	Br	Br	$CH[(CH_2)_5CH_3]CO_2H$	O	-1.6383	2	2	3
67	CH_3	CH_3	$CH(CH_2Ph)CO_2H(R)$	O	-1.1308	2	2	1
68		H	$CH(CH_2Ph)CO_2H(S)$	O	-1.2596	2	2	3

Table 1. (Continued)

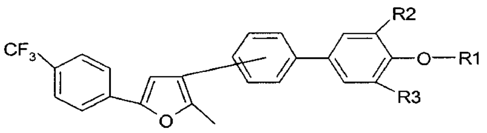
Comp.						Two Class Problem		three Cls
No.	R1	R2	R3	X	Log IC_{50}	TRUE Class	Pred.[a] Class	Pred.[b] Class
69		H	CH_2CO_2H	O	-0.7696	2	2	1
70[c]		H	CH_2CH_2Ph	O	-1.0862	2	2	2
71		H	CH_2CH_2Ph	O	-0.8539	2	2	1
72[d]		H	H	O	-0.0362	1	1	1
73		H	H	O	-0.3372	1	1	1
74		H	H	O	-0.7959	2	2	1
75[c]		CH_2CO_2H	4-OCH_3-Ph 4-OCH_3-Ph		-0.3188	2	1	3
76[c]		CH_2CO_2H	4-OCH_3-Ph 4-OCH_3-Ph		-1.5086	2	2	3

2-Butyl Benzofuran Biphenyls



Comp.						Two Class Problem		Three Cls
No.	R1	R2	R3	X	Log IC_{50}	TRUE Class	Pred.[a] Class	Pred.[b] Class
77	H	H	H	CH_2	0.0755	1	1	1
78[d]	H	H	H	CH(OH)	-0.6383	1	1	1
79	H	Br	Br	CH(OH)	0.1461	1	1	1
80	CH_2CO_2H	H	H	CH_2	0.0607	1	1	1
81	CH_2CO_2H	H	H	CH(OH)	-0.2676	1	1	1
82		H	H	CH_2	-0.2924	1	1	1

Substituted Oxazole Biphenyls



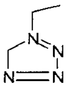
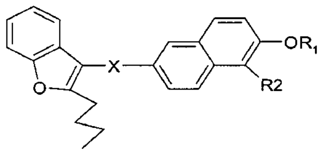
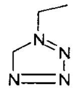
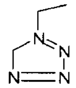
Comp.						Two Class Problem		Three Cls
No.	R1	R2	R3	P.O.A.	Log IC_{50}	TRUE Class	Pred.[a] Class	Pred.[b] Class
83	CH_2CO_2H	H	H	4'	-0.0969	1	1	1
84[c]	$CH(CH_2Ph)CO_2H$	H	H	4'	0.1139	1	1	1
85[d]		H	H	4'	-0.0458	1	1	1

Table 1. (Continued)

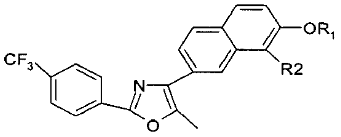
Comp.						Two Class Problem		Three Cls
No.	R1	R2	R3	X	Log IC_{50}	TRUE	Pred.[a]	Pred.[b]
						Class	Class	Class
86	$CH(CH_2Ph)CO_2H$	H	H	3'	0.2041	1	1	1
87[c]	H	Br	Br	4'	-0.1871	1	1	1
88	CH_2CO_2H	Br	Br	4'	-0.3279	1	1	1
89	$CH(CH_2Ph)CO_2H$	Br	Br	4'	-0.8861	2	1	1

2-Butyl benzofuran Biphenyls



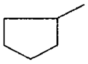
Comp.					Two Class Problem		Three Cls
No.	R1	R2	X	Log IC_{50}	TRUE	Pred.[a]	Pred.[b]
					Class	Class	Class
90	H	H	CH_2	0.1139	1	1	1
91	H	H	CH(OH)	0.0414	1	1	1
92	H	Br	CH(OH)	-0.3188	1	1	1
93	H	Br	CH_2	-0.4815	1	1	1
94	H	I	CH_2	-0.4202	1	1	1
95	CH_2CO_2H	Br	CH_2	0.1461	1	1	1
96[d]	$CH(CH_2Ph)CO_2H$	Br	CH_2	-0.4318	1.	1	1
97	$CH(CH_2Ph)CO_2H$	Br	CO	0.0792	1	2	1
98	$CH(CH_2Ph)CO_2H$	I	CH_2	-0.4949	1	1	1
99		Br	CH_2	-0.1549	1	1	1
100[c]		Br	CO	0.0414	1	1	1

Substituted Oxazole Napthalenes

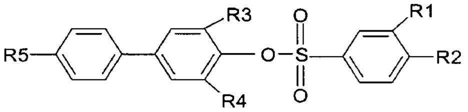


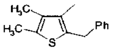
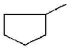
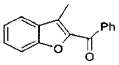
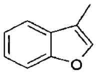
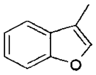
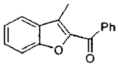
Comp.							Two Class Problem		three Cls
No.	R1			R2		Log IC_{50}	TRUE	Pred.[a]	Pred.[b]
							Class	Class	Class
101	CH_2CO_2H			Br		0.1139	1	1	1
102	H	CO_2H	H	H	O	-1.1249	2	2	3
103	CO_2H		H	H	O	-0.9747	2	2	3
104[c]	OH	CO_2H	H	H	O	-1.4089	2	2	3
105	CO_2H	OH	H	H	O	-1.5850	2	2	3

Table 1. (Continued)

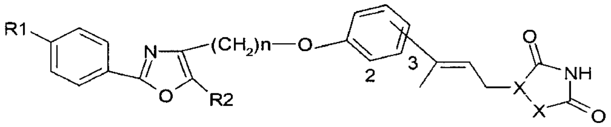
Comp.							Two Class Problem		Three Cls
No.	R1	R2	R3	R4	X	Log IC_{50}	TRUE Class	Pred.[a] Class	Pred.[b] Class
106	OH	CO_2H	CH_3	CH_3	O	-1.4685	2	2	3
107	OH	CO_2H	NO_2	H	O	-1.5376	2	2	3
108	OH	CO_2H		H	O	-1.5528	2	2	3
109[d]	OH	CO_2H	H	H	S	-1.6198	2	2	3
110	OH	CO_2H	Br	H	S	-1.5528	2	2	3
111	OH	CO_2H	Br	Br	S	-1.5229	2	2	3

Sulfono Biphenyls



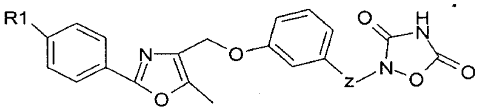
Comp.							Two Class Problem		three Cls
No.	R1	R2	R3	R4	X	Log IC_{50}	TRUE Class	Pred.[a] Class	Pred.[b] Class
112	OH	CO_2H	H	H		-1.4949	2	2	3
113	OH	CO_2H		H		-0.4510	1	2	2
114	OH	CO_2H	H	H		0.0645	1	1	1
115	OAc	CO_2H	H	H		-1.3979	2	2	2
116	OH	CO_2H	NO_2	H		-0.7496	2	2	1

Oxazole Azolidinediones



Comp.								Two Class Problem		Three Cls
No.	R1	R2	N	P.O.A.	X	Y	Log IC_{50}	TRUE Class	Pred.[a] Class	Pred.[b] Class
117	CF_3	Ph	1	3	N	O	-0.5229	1	1	1

Oxazole Azolidinediones



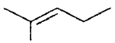
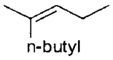
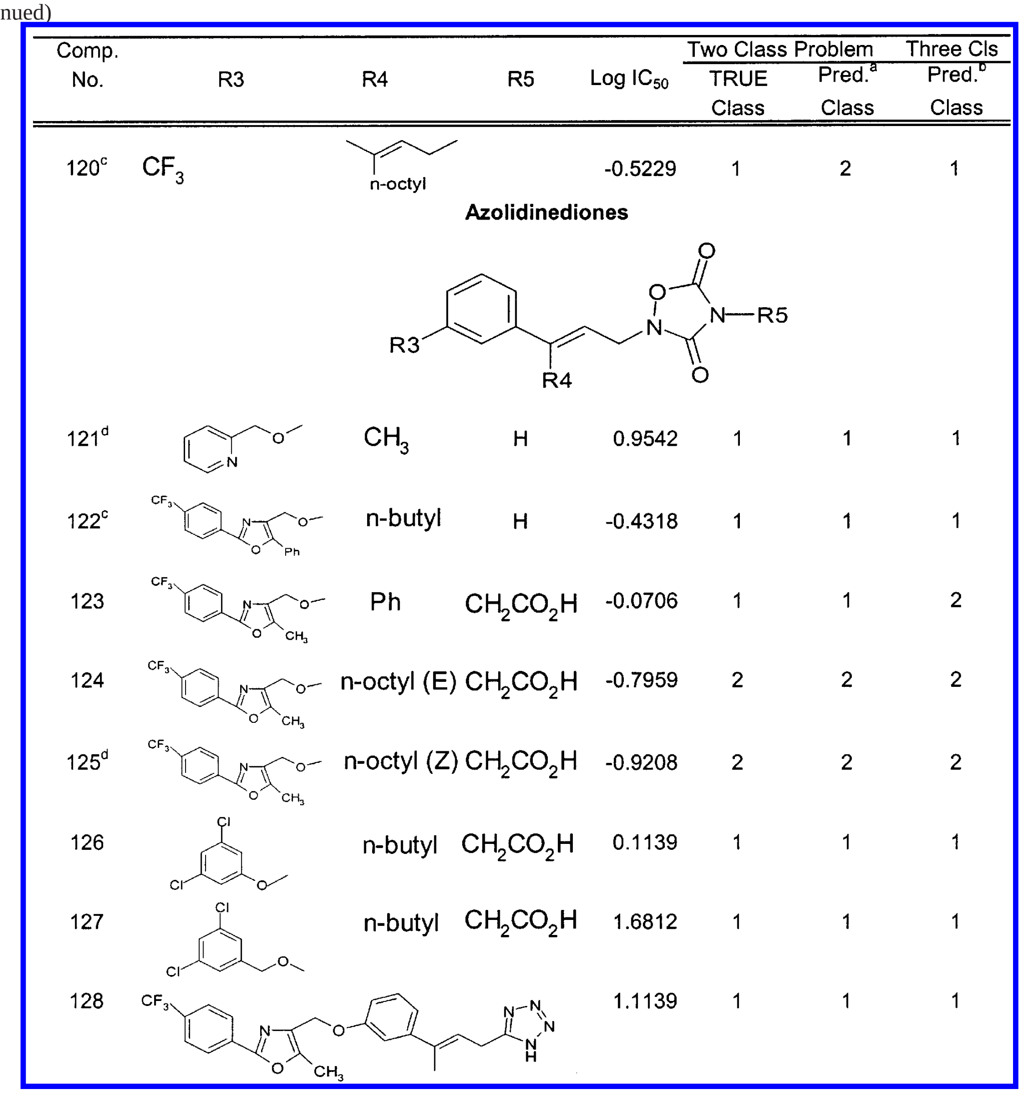
Comp.				Two Class Problem		Three Cls
No.	R1	Z	Log IC_{50}	TRUE Class	Pred.[a] Class	Pred.[b] Class
118[d]	CF_3	 n-propyl	0.2788	1	1	1
119	CF_3	 n-butyl	0.1461	1	1	1

Table 1. (Continued)


n-octyl
R3
R4
O
O
N
N—R5
O
N
CF3
Ph
CH3
Cl
N–N
N
N
H

Comp. No.	R3	R4	R5	Log IC_{50}	Two Class Problem TRUE Class	Pred.[a] Class	Three Cls Pred.[b] Class
120[c]	CF_3	n-octyl		-0.5229	1	2	1
		Azolidinediones					
121[d]		CH_3	H	0.9542	1	1	1
122[c]		n-butyl	H	-0.4318	1	1	1
123		Ph	CH_2CO_2H	-0.0706	1	1	2
124		n-octyl (E)	CH_2CO_2H	-0.7959	2	2	2
125[d]		n-octyl (Z)	CH_2CO_2H	-0.9208	2	2	2
126		n-butyl	CH_2CO_2H	0.1139	1	1	1
127		n-butyl	CH_2CO_2H	1.6812	1	1	1
128				1.1139	1	1	1

[a] All values calculated from best two-classes model. [b] All values calculated from best three-class model. [c] Prediction set compounds. [d] cvset compounds.

Table 2. Distribution of Compounds into TSETs and PSETs for Two-Class and Three-Class Problem

	TSET	PSET
Two-Class Problem		
inactive (log $IC_{50} > -0.7$)	61	8
active (log $IC_{50} \leq -0.7$)	53	6
Three-Class Problem		
inactive (log $IC_{50} > -0.7$)	61	8
moderate (log $IC_{50} \leq -0.7$ to -1.2)	28	2
active (log $IC_{50} \leq -1.2$)	25	4

to a reasonable level. In practice the ratio of descriptors for compounds to the number of TSET observations used was less than or equal to 0.6. This was carried out using 114 TSET observations for both the topological pool and the all-descriptor pool. Any descriptor containing identical values for 90% or more of the TSET observations was eliminated. Additionally, pairwise correlations were calculated for all descriptors. One of any two descriptors with a correlation above 0.9 was eliminated. These two steps reduced the topological pool of 75 descriptors to 37 descriptors and the all-descriptor pool of 94 descriptors to 38 descriptors. Flexibility descriptors were present in both pools. These were acceptable levels as the ratio of descriptors to TSET observations was well below 0.6 for all cases investigated. Models based on only topological descriptors were evaluated first as they offer an advantage of not requiring geometry-optimized structures. Then models were developed using all descriptors.

Model Development and Validation. The two reduced descriptor pools were screened using genetic algorithm[37,38] (GA) evolutionary optimization. The GA feature selection routines were written in-house. Different subsets of descriptors were evaluated to see if they could develop classifiers to determine the PTP 1B inhibitory activity. Models were formed using k-nearest neighbor analysis (k-NN), linear discriminant analysis (LDA),[39,40] and radial basic function neural networks (RBFNN).

k-NN is a nonparametric classification technique[41,42] that takes an unknown input pattern and classifies it to the class of the majority among its k-nearest neighbors in the training on the basis of Euclidian distance metric. It is effective when probabilities of the feature variables are unknown. As the number of descriptors used in the model building increases the probability of finding models due to chance correlation increases, so model sizes ranging from 3 to 10 descriptors were evaluated. For each model size GA was used to search the descriptor space for the subset of descriptors that produce the lowest COST function. COST function is further discussed later in the paper. Best models were chosen on the basis of lowest COST function and least number of descriptors. Once selected the optimum model was used to classify compounds in the PSET to verify generalization ability of the model.

LDA is a supervised classification technique that maximizes separation between class means in descriptor space, relative to standard deviation. Discriminants were generated using the TSET compounds. Again, model sizes ranging from 3 to 10 descriptors were investigated. For each model size, GA was used to search the descriptor space for the subset of descriptors that produce the lowest COST function. Once selected the optimum model was used to classify compounds in the PSET to verify the generalization ability of the model.

The RBFNN classifier was developed in-house, in an effort to introduce nonlinearity in the classifiers. Thus, it would provide a potential advantage over the linear classifiers. This classifier has one hidden layer. In the training phase the network parameters were determined, and subsequently the output layer was adjusted. Generally implementing RBFNN classifier is a complex task.[43] Usually these classifiers need the user to input a number of parameters. In this classifier the user input is minimized. Several radial basis functions such as eqs 1−5 were explored

$$h(r) = \mathrm{e}^{-r^2/2\,\sigma^2} \tag{1}$$

$$h(r) = (r^2 + \sigma^2)^{1/2} \tag{2}$$

$$h(r) = (r^2 + \sigma^2)^{-1/2} \tag{3}$$

$$h(r) = r^2 * \log r \tag{4}$$

$$h(r) = r^2/\sigma^2 * \log r/\sigma \tag{5}$$

where h is the radial basis function and σ is the spread parameter. GA routines were developed and were used to search the descriptor space for the subset of descriptors that produce the lowest COST function. Model sizes ranging from 3 to 10 descriptors were investigated.

The COST function computed for each descriptor subset needs to generalize well, and it should not favor one class at the cost of the other. A leave-N-out cross validation was used in this GA routine with N as 10% of the TSET compounds. This tested the ability of each model to generalize the compounds left out. Thus, the models that could only predict the TSET compounds were avoided. In addition models were avoided which reduced classification overall by increasing misclassifications for a particular class, as this would affect data sets where there is an uneven distribution in the number of compounds in each class.

The computations for this work were performed on a DEC 3000 AXP Model 500 workstation. Those calculations involving HyperChem[24] were performed on a Pentium PC.

QSAR. Initially the 12 compounds beyond our software limits were set aside as an exclusion set. This reduced the PSET form 14 to 12 compounds. TSET changed from 114 to 104 compounds. For the development of a multiple linear regression model, the TSET included all of the 104 compounds. For the generation of nonlinear CNN models, the training set was further subdivided into a training set containing 92 compounds and a cross-validation set (CVSET) containing 12 compounds. Table 1 lists these compounds and their experimental log(IC_{50}) values.

About 90 topological descriptors generated previously in the classification study were reduced by using objective feature selection process to a reduced pool of 37 descriptors. At this stage, more than 200 descriptors were generated using geometry optimized structures. They included topological and geometric as well as electronic descriptors. Descriptors such CPSA, volume, and surface area, which could not be calculated for the classification study, due to software limitations were also part of this pool. These descriptors were then reduced using the objective feature selection process and vector space descriptor analysis to a reduced pool containing 48 descriptors. Vector space descriptor analysis routine treats the descriptors as multidimensional vectors and uses a stepwise orthogonalization procedure to find a subset of mutually orthogonal descriptors, which further lowers the likelihood of chance correlation. However, this reduced pool of descriptors still contains most of the variance in the data. At first, models based on only topological descriptors were evaluated as they offer an advantage of not requiring geometry-optimized structures. Then models were developed on the basis of all descriptors.

Multiple linear regression models (Type 1 models) that link the property of interest to the structures can be developed using subsets of descriptors selected from the reduced descriptor pool. A simulated annealing[44,45] feature selection routine and a genetic algorithm[46] feature selection routine were used to find good descriptor subsets. Each method performed a directional search of the descriptor space, to determine an optimal subset of descriptors to be used in a linear regression model. The driving force behind each algorithm is the continuous reduction of the root-mean-square (rms) error of log(IC_{50}) value estimation from subset to subset. Subsets of descriptors that give lower rms errors are favored. An interactive regression routine was used to generate statistically valid linear models from these subsets. The linear regression models were explored first for the reduced pool of descriptors. They were validated using t-values to ensure that the model coefficients were not zero and multiple linear correlations to detect multicollinearities. The quality of models was based on rms error. The best model was then used to predict estimated log(IC_{50}) values for the prediction set compounds.

The set of descriptors chosen from the linear model was subsequently submitted to computational neural networks (CNN) to develop a nonlinear (Type 2) model. In this study, a fully connected, three layered, feed forward network was trained using a quasi-Newton optimization algorithm. Detailed discussions of the type of neural network and the training algorithm used in this study have been published previously.[47,48] The set of descriptors chosen for the best linear model were used as input neurons. The number of hidden layer neurons was varied to find the best nonlinear model. The best model was selected on the basis of fewer neurons and lower rms error.

The same procedure was repeated for the reduced pool containing 48 topological, geometric, and electronic descriptors. In addition, the 12 excluded compounds were added to the study, thus using the same TSET and PSET used in the classification. Approximately 80 topological descriptors were calculated for all 128 compounds. However, the flexibility descriptors were not part of this pool as the routine was developed at a later stage. The descriptors were reduced to a 40-descriptor reduced pool, and linear and nonlinear model development was investigated. All the models that were developed in this study were validated with the same external prediction set.

RESULTS AND DISCUSSION

Classification. GA was used to evaluate subsets with three to 10 descriptors. The smallest descriptor subset that produced an acceptably low COST function was selected as optimal. These descriptor values for the TSET compounds were then used to develop the models. Once generated these models were used to classify all the compounds in the TSET and calculate the percent correct values. Then the compounds of the external prediction set were classified, and the percent correct values were calculated for them.

Two-Class Problem. First, the problem of generation of binary models to differentiate actives from inactives was explored. Models were investigated using the reduced pool of 38 topological descriptors, using k-NN and LDA classification techniques. The optimal model formed from this reduced pool by k-NN classifier was a six-descriptor model. Of the 114 TSET compounds, 89.5% were classified correctly. The correct classification rate for the 14 PSET compounds was 85.7%, which clearly demonstrates that the optimum model is capable of classifying compounds not used in the model formation.

A second model was generated to address the same active/inactive problem using the all-descriptor reduced pool. The optimal model chosen using k-NN classifier had five descriptors. It predicted 91.2% of the TSET compounds correctly and 85.7% of the PSET compounds correctly. As this model has slightly better prediction rates for a smaller number of descriptors, this model is superior according to the criteria for model development.

Models were also developed with an LDA classifier using both reduced pools. Results similar to the k-NN results were obtained using LDA for both reduced descriptor pools. These models are not discussed in this paper.

The newly developed RBFNN classifier with the most commonly used Gaussian radial basis function (eq 1) did not yield the best model among all the RBFNN models. The RBFNN classifier developed using the eq 2 yielded the best model. This optimal model was a six-descriptor model. Of the 114 PSET compounds 81% were classified correctly. The classification rate for the CVSET was 81.4% and the 14 compound PSET was 80%. Thus, a reasonable rate of classification was achieved. There is still some room for improvement, and future work will address that question.

The six topological descriptors selected in the best model are shown in Table 3. EMAX1 denotes the maximum atomic electrotopological state value.[30] Electrotopological state values provide information about intermolecular interactions. The MOLC2 represents molecular connectivity corrected for the rings. MDE24 describes the connection information between secondary and quaternary carbons.[32] MOLC5 counts all path lengths of length three. The NC2 descriptor is simply the count of carbons. This is not surprising based on the characterization of the data set. WTPT4 calculates the sum of all path weights starting from oxygen atoms. The two MOLC descriptors with flags two and five give information about degree of branching, which relates to the flexibility of the compounds. Flexibility of the compounds is an important factor influencing the binding at the PTP 1B site. Table 3 also shows the range and average for the inactive and active classes for each of the six descriptors as well as the relative variance for each descriptor. Attention should be drawn to the fact that the average descriptor value for the active compounds is always greater than the average descriptor value for the inactive compounds. EMAX1, MOLC2, and MOLC5 have comparatively low relative variance values; however, they provide useful information as they improve the model prediction rates.

Table 3. Six Topological Descriptors Defining the Optimal Two-Class Model

	range		average		
descriptor[a]	active	inactive	active	inactive	rel. var[b]
EMAX1	11.1−15.7	6.20−15.4	13.1	12.4	0.22
MOLC2	9.36−16.9	8.11−16.3	13.5	11.3	0.34
MDE24	0.0−56.0	0.0−52.1	15.9	13.6	7.38
MOLC5	5.83−10.1	3.59−9.35	7.91	6.28	0.25
NC2	24.0−43.0	18.0−43.0	35.7	29.3	0.99
WTPT4	7.64−25.7	2.37−20.9	12.9	10.1	1.65

[a] Explanations: EMAX1, maximum E-state value;[30] MOLC2, molecular connectivity corrected for rings;[50] MDE24, distance edge between S−Q carbons; [32] MOLC5, path length of three;[50] NC2, number of carbons; WTPT4, sum of all path weights starting from oxygen.[41] [b] Relative variance is the variance divided by the mean for a descriptor using all compounds for both the classes.

This model predicted 12 out of 14 PSET compounds correctly. All the active compounds were correctly classified by this model. Compounds 7 and 13 were the two inactive compounds that were misclassified. Compound 7 was the only compound in the data set that had a primary carbon attached to the oxygen on the biphenyl ring as seen from Table 1. So this misclassification may be due to lack of exposure in the training phase to primary carbon substitution on the biphenyl ring. Compound 13 is the only compound with a tertiary carbon substituted on the oxygen connected to the biphenyl ring. This compound also had no representation in the training phase.

To ensure that the results were not due to chance, Monte Carlo experiments were conducted in which models were generated after scrambling of class labels. These results were close to the random assignments. The result clearly demonstrated that the predictive ability of the six-descriptor k-NN model was very unlikely to have been due to chance.

Three-Class Problem. Models were generated to differentiate active, moderately active, and inactive compounds, using the distribution shown in Table 2. When the topological descriptor reduced pool was considered, the best model generated using the k-NN classifier contained six descriptors. It predicted 78.9% of the TSET compounds correctly and 78.6% of the PSET compounds correctly. The optimal model formed from the same reduced pool by the LDA classifier was a four-descriptor model. Of the 114 TSET compounds, 71.9% were classified correctly and out of 14 PSET

Table 4. Seven Topological Descriptors Defining the Optimal Three-Class Model

	range		average		
descriptor[a]	active	inactive	active	inactive	rel. var[b]
ENEG	4.04−5.22	4.09−5.29	4.46	4.57	0.02
PEND6	0.00−1.46	4.09−5.29	50.8	92.8	607.7
NBND	32.0−56.0	27.0−56.0	47.3	39.1	0.54
MOLC5	5.83−10.1	3.59−9.35	7.91	6.28	1.12
HOMO	−8.86 to −7.15	−9.06 to −7.6	−7.99	−8.33	0.02
EAVG1	0.559−1.01	0.40−0.97	0.79	0.66	0.03
EAVG2	7.12−10.5	4.26−10.8	8.58	8.52	0.15

[a] Explanations: ENEG, electronegativity (0.5(homo+lumo)); PEND6, vertices of all pendant halogens;[31] NBND, number of double bonds; MOLC5, path length of three;[50] HOMO, highest occupied molecular orbital; EAVG1, average E-state over all heavy atoms;[30] EVG2, average E-state over all heteroatoms.[30] [b] Relative variance is the variance divided by the mean for a descriptor using all compounds for both the classes.

Table 5. Confusion Matrix for the TSET and PSET Compounds Using the Optimal Seven Descriptor Model for the Three-Class Problem

	predicted class			
actual class	inactive	moderate	active	% correct
	a. TSET Compounds			
inactive	56	4	1	91.6
moderate	8	15	5	53.6
active	0	2	23	90.0
	b. PSET Compounds			
inactive	7	0	1	87.5
moderate	0	1	1	50.0
active	0	0	4	100.0

compounds 78.6% were classified correctly. These classification rates were lower than those obtained for active/inactive situation, but that is expected, as this three-class problem is considerably more difficult.

A second model was generated to address the same active/moderately active/inactive problem using the all-descriptor reduced pool. This descriptor pool contained an additional five flexibility descriptors. More than 50% of the top five models in descriptor subsets between three and 10 contained at least one flexibility descriptor. The optimal model has seven descriptors, which was developed using the k-NN classifier. It predicted 82.5% of the TSET compounds correctly and 85.4% of the PSET compounds correctly. The LDA classifier led to a five-descriptor model using the same reduced pool. It predicted 79.0% of the TSET compounds correctly and 78.6% of the PSET compounds correctly. These classification rates are much higher than the expected random results. The seven descriptor model was considered optimal.

The seven descriptors selected in the best model are shown in Table 4. Four of the seven descriptors are energy related in this all-descriptor pool. MOLC5 and NBND values indicate that the branching favors the increase in activity. The average value for PEND6 is less for the actives, which seems to indicate that the less number of halogens improves the activity. ENEG, HOMO, and EAVG1 have comparatively low relative variance values; however, they provide useful information as they improve the model prediction rates.

Table 5a shows the confusion matrix for the TSET compounds using the seven-descriptor model developed from all-descriptor reduced pool. Also shown are the correct classification rates for all three classes using this model. The table demonstrates that the cost function found descriptor subsets that are not biased for or against any one class for the active and inactive TSET compounds. As the moderately active class is very small, the prediction models seemed to misclassify more moderate compounds to increase the accuracy for the active and inactive classes.

The classification rate for the PSET compounds is close to or better than that obtained for the TSET compounds, as is clearly seen from Table 5b. Accuracy is good for the active and inactive classes but poor for the moderate class. This may be due to the fact that there were only two moderate compounds in the PSET, out of which one was classified incorrectly. In an effort to keep the prediction set completely external for all the work done on this data set, the distribution was not changed to make the PSET reasonably represented in all three classes. This added more difficulty to the problem at hand. The poor classification of moderates may be an unavoidable consequence of uneven class populations. It is also important to note that despite these problems the active compounds had a classification rate of 100%.

No active compounds were misclassified in the PSET. There were two compounds misclassified in the PSET, one in the moderate class and one in the inactive class. Again the same two compounds (7 and 13) were misclassified. It seems that with the random PSET generation the only primary as well as the only tertiary compound substituted on the oxygen attached to the biphenyl ring ended up in the PSET. This left no representation in the training phase, leading to the misclassification.

Monte Carlo experiments were conducted on this data set in which models were generated after scrambling of class labels. These results were close to the random assignments. The result clearly demonstrated that the predictive ability of the seven descriptor model is very unlikely to have been due to chance.

QSAR. Finally, a series of QSAR models were generated to quantitatively predict the log(IC_{50}) values. Many potential linear models were investigated. The multiple linear regression model containing smallest subset of descriptors with most favorable *F* values, multiple correlation coefficient, and rms error was chosen for further investigation. The best Type 1 linear model found using the topological descriptor pool consisted of eight descriptors. This model encoded information about the topology of the inhibitors. The TSET had rms error of 0.321 log units and a multiple correlation coefficient of 0.859. In validation of this model, the external prediction set had an rms error of 0.447 log units. This linear model showed a substantial amount of scatter. It was evident that the relationship between structure and log(IC_{50}) was not completely linear. This led to the exploration of nonlinear relationships between the structures and log(IC_{50}).

The descriptors in the best topological linear model were submitted as inputs to a CNN. The best architecture found after varying the number of hidden neurons was 8−2−1. This architecture had a ratio of adjustable parameters to number of observations above 2.0. The TSET, CVSET, and PSET rms errors were 0.332, 0.333, and 0.385 log units, respectively. The PSET showed a significant improvement over the linear model as PSET error improved by about 17%.

The best Type 1 linear model found using the all-descriptor pool consisted of eight descriptors. Table 6 shows these

Table 6. Descriptors Used in Linear Type 1 Model and Nonlinear Type 2 Model for PTP 1B Inhibitors (116 Compounds)

descriptor	coeff[a]	error est[a]	range	explanation[b]
MOLC3	−0.834	0.133	6.90−14.0	path 4 molecular connectivity
MOLC6	−0.119	0.135	12.2−32.0	valence and ring corrected conn.
MOLC9	−2.49	0.081	1.12−1.83	path length of 2
NO3	0.442	0.136	1−10	number of oxygens
NDB13	0.466	0.113	0−4	number of double bonds
FNSA3	23.5	0.146	−0.04−0.17	fractional negative surface area
RNCG1	−16.5	0.095	0.06−0.18	relative negative charge
SCAA2	0.154	0.114	−12.1−12.9	surface area per acceptor atom

[a] Values representing linear model. [b] MOLC3, molecular connectivity corrected for valence and rings;[50] MOLC6, molecular connectivity for path length of 4;[50] MOLC9, path length of 2; [50] NO3, number of oxygens; NDB, number of double bonds; FNSA3, fractional negative surface area;[36] RNCG1, relative negative charge;[36] SCAA2, surface area × charge of hydrogen bond acceptor atoms/number of hydrogen bond acceptor atoms.[51]

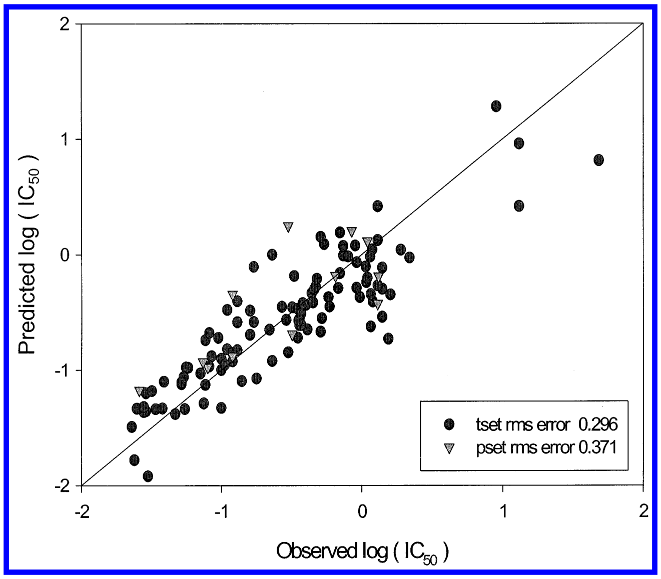


Figure 1. Predicted vs experimental log(IC_{50}) values for Type 1 linear model generated using all descriptors.

descriptors with their coefficients and errors. Half of the descriptors in this model encoded information about hydrogen bonding and charge, which is not surprising. The polar surface area descriptors such as FNSA and RNCG incorporated the effect of charge and size in the above model. The SCAA descriptors influenced the contribution of donor nitrogen and acceptor hydrogen atoms. The other half of the descriptors encoded the topology of the inhibitors. The three MOLC descriptors again coded information about branching. The very simple number of double bonds and number of oxygens were a little unusual; however, they did make a positive contribution to the model.

A plot of experimental log(IC_{50}) values vs predicted log(IC_{50}) values for the Type 1 linear model derived from the all-descriptor pool is shown in Figure 1. The TSET had an rms error of 0.296 log units and a multiple correlation coefficient of 0.889. In validation of this model, the external prediction set had an rms error of 0.371 log units. This linear model shows a substantial improvement over the topological model. Therefore the inclusion of CPSA and hydrogen bonding descriptors improved the model considerably. Even though this is a better model, it is evident that the relationship between structure and log(IC_{50}) is not completely linear. So further exploration of the nonlinear relationships between the structure and the property is necessary.

The descriptors in the best all descriptor linear model were submitted as inputs to CNN. The best architecture found after varying the number of hidden neurons was 8−3−1. This architecture had a ratio of adjustable parameters to a number of observations above 2.0. The TSET, CVSET, and PSET rms errors were 0.279, 0.287, and 0.302 log units, respectively. The PSET showed a significant improvement over the linear model as PSET error improved by about 12%. Figure 2 shows the plot of experimental vs predicted log(IC_{50}) values by the Type 2 nonlinear CNN model. As expected this nonlinear model was considerably better than the topological one. There was very little scatter as is seen from Figure 2.

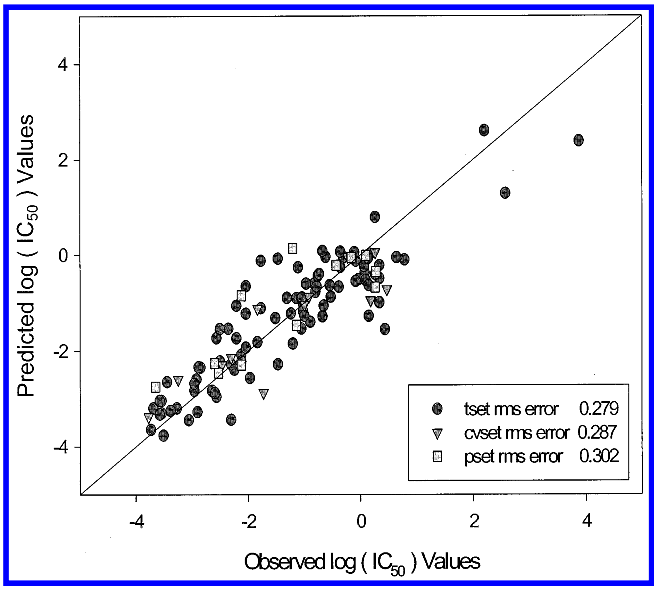


Figure 2. Predicted vs experimental log(IC_{50}) values for Type 2 nonlinear CNN model generated using all descriptors.

In addition, a Type 1 linear model was developed using the 128 compounds and topological descriptor reduced pool. The best Type1 linear model found using this all descriptor pool again consisted of eight descriptors. The TSET had an rms error of 0.321 log units and a multiple correlation coefficient of 0.889. In validation of this model, the external prediction set had an rms error of 0.410 log units. This linear model is comparable to the other topological model without the 12 compounds beyond the software limit. As the two topological models, with and without the 12 compounds beyond the software limit, were similar the descriptors from this linear model were not submitted to a neural network.

SUMMARY AND CONCLUSIONS

Classification. Computational models were developed for a small data set of heterocycles with substitution on 10

Table 7. Summary of Prediction Results for All Optimal k-NN Models Investigated

no. of classes	descriptors considered	no. of descriptors	% correct	
			TSET	PSET
2	topological	6	89.5	85.7
2	all	5	91.2	85.7
3	topological	6	78.9	78.6
3	all	7	82.5	85.4

different moieties, using a topological descriptor pool as well as topological, electronic, and flexibility descriptor pool using different classification methods. Despite a very small sized PSET (4 active, 2 moderately active, 8 inactive) no false negatives were found for PSET predictions using all optimal models. In both two-class and three-class problems over 80% classification rate was achieved. The newly developed flexibility descriptor proved useful in the current study.

Successful model development on the basis of topological descriptors for the two-class problem is beneficial as a screening method for large databases of compounds as topological descriptor calculations are not computationally intensive. Another advantage of some of these descriptors is that they do not need accurate 3-D geometries and conformational specificities. This model correctly classified 12 out of 14 PSET compounds correctly. In addition all the active PSET compounds (6 out of 6) were classified correctly. The two compounds that were misclassified had no representation in the training phase.

The newly developed RBFNN classifier yielded reasonable results of 80% correct classifications for the two-class problem. Further improvements in this classifier are being investigated.

For a three-class problem the topological descriptors yielded significantly poorer models. Addition of energy and flexibility descriptors improved the model classification rate from about 70% to about mid 80%. These models were far superior to the ones developed by using only topological descriptors. Again this model correctly classified 12 out of 14 PSET compounds correctly. In addition all the active PSET compounds (4 out of 4) were classified correctly. The two compounds that were misclassified (7 and 13) were the same compounds that were misclassified in the two-class problem. So overall the model does not seem to extrapolate as well. However considering the small sizes this is a good classification rate.

Overall good classification accuracies were obtained for this data set using these descriptors. Flexibility was selected in about 40% of all top models in the GA search. These models were developed from an external prediction set, which was not exposed during the process of model development and still yielded good classification rates for the two-class (active, inactive) as well as three-class (active, moderate, inactive) problem. In addition to good classification rates the models had no false negatives.

A summary of results obtained using k-NN is shown in Table 7. Prediction errors produced using models with topological descriptors were consistently higher than the ones produced using all descriptor pool. The addition of quantum chemical descriptors in the descriptor pool proved to improve models, as either the error decreased or the number of descriptors needed decreased. As models developed with topological descriptors are computationally cost efficient and are invariant with regard to the geometry of structures they are still useful for fast estimates.

QSAR. A series of models were developed using QSAR methodology. The models clearly demonstrate a connection between structure and inhibitory concentrations of PTP 1B inhibitors. These are the first models that predict IC_{50} values for PTP 1B inhibitors of several variants of azolidenediones and benzofuran benzothiophene biphenyls. The structural information of PTP 1B inhibitors is numerically encoded as molecular descriptors. Objective feature selection and vector space analysis led to development of linear models. Theses were further refined to develop nonlinear models. A nonlinear feature selection routine, which combined the genetic algorithm with a neural network fitness evaluator, was used to develop CNN models. All models were validated using the same external prediction set.

This study confirms that IC_{50} values can be predicted on the basis of molecular structure alone, without the inclusion of any experimentally derived data such as partition coefficients. The regression and CNN models can be applied to prediction of inhibitory activities of compounds that are not present in the data set used in this study as long as they are structurally similar. The predictive power of these models could be useful in cases where biological assays are necessary to measure the activities of different pharmacophores.

REFERENCES AND NOTES

(1) http://www.albmolecular.com/features/tekreps/vol04/no44.

(2) DeFronzo, R. A.; Bonadonna, R. C.; Ferrannini, E. Pathogenesis of NIDDM. *Diabetes Care* **1992**, *15*, 318−368.

(3) Reaven, G. Role of Insulin Resistance in Human Disease. *Diabetes* **1988**, *37*, 1595−1607.

(4) Goldman, J. M. Oral Hypoglycemic agents: An Update of Sulfonylureas. *Drugs Today* **1989**, *25*, 689−695.

(5) Kolterman, O. G.; Prince, M. J.; Olefsky, J. M. Insulin Resistance in Non-Insulin Dependent Diabetes Mellitus. Impact of sulfonylureas agents in vivo and in vitro. *Am. J. Med.* **1983**, *74* (Suppl. 1A), 82−101.

(6) Ferrannini, E. The Insulin Resistance Syndrome. *Curr. Opin. Nephrol. Hypertens.* **1992**, *1*, 291−298.

(7) Jackson, M. D.; Denu, J. M. Molecular Reactions of Protein Phosphatases − Insights from Structure and Chemistry. *Chem. Rev.* **2001**, *101*, 2313−2340.

(8) Zhan, X. L.; Wishart, M. J.; Guan, K. L. Nonreceptor Tyrosine Phosphatases in Cellular Signaling; Regulation of Mitogen − Activated Protein Kinases. *Chem Rev.* **2001**, *101*, 2477−2496.

(9) Hunter, T. Protein Kinases and Phosphatases: the yin and yang of Protein Phosphorilation and Signalling. *Cell* **1995**, *80*, 225−236.

(10) Tonks, N. K.; Neel, B. G. From Forms to Function: Signaling by Protein Tyrosine Phosphatases. *Cell* **1996**, *87*, 365−368.

(11) Elchebly, M.; Payette, P.; Michaliszyn, E.; Cromlish, W.; Collins, S.; Loy, A. L.; Normandin, D.; Cheng, A.; Himms-Hagen, J.; Chan, C. C.; Ramachandran, C.; Gresser, M. J.; Tramblay, M. L.; Kennedy, B. P. Increased Insulin Sensitivity and obesity resistance in mice lacking the protein tyrosine phosphatase- 1B gene. *Science* **1999**, *283*, 1544−1548.

(12) Desmarais, S.; Friesen, R. W.; Zamboni, R.; Ramachandran, C. [Difluro(phosphono)methyl]phenylalanine-containing peptide inhibitors of protein tyrosine phosphatases. *Biochem. J.* **1999**, *337*, 219−223.

(13) Iversen, L. F.; Andersen, H. S.; Moller, K. B.; Olsen, O. H.; Peters, G. H.; Branner, S.; Mortensen, S. B.; Hansen, T. K.; Lau, J.; Ge, Y.; Holsworth, D. D.; Newman, M. J.; Moller, N. P. H. Steric Hindrance as a Basis for Structure-Based Design of Selective Inhibitors of Protein-Tyrosine Phosphatases. *Biochemistry* **2001**, *40*, 14812−14820.

(14) Sarmiento, M.; Wu, L.; Keng, Y. F.; Song, L.; Luo, Z.; Huang, Z.; Wu, G. Z.; Yuan, A. K.; Zhang, Z. Y. Structure-Based Discovery of Small Molecule Inhibitors Targeted to Protein Tyrosine Phosphatase 1B. *J. Med. Chem.* **2000**, *43*, 146−155.

(15) Malamas, M. S.; Serdy, J.; Moxham, C.; Katz, A.; Xu, W.; McDevitt, R.; Adebayo, F. O.; Sawicki, D. R.; Seestaller, L.; Sullivan, D.; Taylor, J. R. Novel Benzofuran and Benzothiophene Biphenyls as Inhibitors of Protein Tyrosine Phosphatase 1B with Antihyperglycemic Properties. *J. Med. Chem.* **2000**, *43*, 1293−1310.

(16) Malamas, M. S.; Serdy, J.; Gunawan, I.; Mihan, B.; Sawicki, D. R.; Seestaller, L.; Sullivan, D.; Flam, B. R. New Azolidinediones as Inhibitors of Protein Tyrosine Phosphatase 1B with Antihyperglycemic Properties. *J. Med. Chem*. **2000**, *43*, 995−1010.

(17) Bradford, D. P.; Keller, J. C.; Flint, A. J.; Tonks, N. K. Purification and Crystallization of the Catalytic Domain of Human Protein Tyrosine Phosphatase 1B Expressed in *Escherichia coli*. *J. Mol. Biol*. **1992**, *239*, 726−730.

(18) Ramachandran, C.; Aebersold, R.; Tonks, N. K.; Pot, D. A. Sequential dephosphorylation of a Multyply Phosphorylated Insulin Receptor Peptide by Protein Tyrosine Phosphatases. *Biochemistry* **1992**, *31*, 4232−4238.

(19) Lanzetta, P. A.; Alverez, L. J.; Reinach, P. S.; Candia, O. A. An Improved Assay for Nanomolar Amounts of Inorganic Phosphate. *Anal. Biochem*. **1979**, *100*, 95−97.

(20) Wessel, M. D.; Jurs, P. C.; Tolan, J. W.; Muskal, S. M. Prediction of Human Intestinal absorption of Drug Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 726−735.

(21) Patankar, S. J.; Jurs, P. C. Prediction of IC_{50} Values for ACAT Inhibitors from Molecular Structure. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 716−723.

(22) Johnson, S. R.; Jurs, P. C. Prediction of Acute Mammalian Toxicity from Molecular Structure for a diverse Set of Substituted Anilines Using Regression Analysis and Computational Neural Networks. In *Computer-Assisted Lead Finding and Optimization*; van de Waterbeemd, H., Testa B., Folkers, G., Eds.; Verlag Helvetica Chimica Acta: Basel, 1997; pp 29−48.

(23) Johnson, S. R.; Jurs, P. C. Prediction of the Clearing Temperatures of a Series of Liquid Crystals from Molecular Structure. *Chem. Mater.* **1999**, *11*, 1007−1023.

(24) Hypercube Inc. Waterloo, OH.

(25) Stewart, J. P. P. MOPAC 6.0; Quantum Chemistry Program Exchange, Indiana University, Bloomsburg, IN, Program 455.

(26) Stewart, J. P. P. MOPAC− A Semiempirical Molecular-Orbital Package. *J. Comput.-Aided Mol. Des*. **1990**, *4*, 1−45.

(27) Stuper, A. J.; Brugger, w. E.; Jurs, P. C. *Computer-Assisted Studies of* chemical *Structure and Biological Function*; Wiley-Interscience: New York, 1979.

(28) Jurs, P. C.; Chou, T. J.; Yuan, M. *In Computer-Assisted Drug Design*; Olsen, E. C., Christoffersen, R. E., Eds.; American Chemical Society: Washington, DC, 1979; pp 103−129.

(29) Johnson, S. R. Ph.D. Thesis, Penn State University, 1999.

(30) Kier, L. B.; Hall, L. H. The E-State as an Extended Free Valence. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 548−552.

(31) Madan, A. K.; Gupta, S.; Singh, M. Superpendentic Index: A Novel Highly Discriminating Topological Descriptor for Predicting Biological Activity. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 272−277.

(32) Cao, C. Distance-Edge Topological Index-Research on Structure−Property Relationships of Alkanes. *Huaxue Tongbao* **1996**, *22*, 1238−1244.

(33) Balaban, A. T. Highly Discriminating Distance-Based Topological Index. *Chem. Phys. Lett*. **1982**, *89*, 399.

(34) Bondi, A. van der Waals Volumes and Radii. *J. Phys. Chem.* **1964**, *68*, 441−451.

(35) Stouch, T. R.; Jurs, P. C. A Simple Method for Representation, Quantification, and Comparison of the Volumes and Shapes of Chemical Compounds. *J. Chem. Inf. Comput. Sci*. **1986**, *26*, 4−12.

(36) Stanton, D. T.; Jurs, P. C. Development and Use of Charged Surface Area in Computer-Assisted Quantitative Structure−Property Relationship Studies. *Anal. Chem.* **1990**, *62*, 2323−2329.

(37) Luke, B. T. Evolutionary Programming Applied to the Development of Quantitative Structure−Activity Relationships and Quantitative Structure−Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1279−1287.

(38) Kimura, T.; Hasegawa, K.; Fanatsu, K. GA Strategy for Variable Selection in QSAR Studies: GA-Based Region Selection for CoMFA Modeling. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 276−282.

(39) Huberty, C. J. *Applied Discriminant Analysis*; John Wiley & Sons: New York, 1994.

(40) Johnson, R. A.; Wichern, D. W. *Applied Multivariate Statistical analysis;* Prentice Hall: Englewood Cliffs, NJ, 1982.

(41) Duda, R. O.; Hart, P. E. *Pattern Classification and Scene Analysis;* John Wiley & Sons: New York, 1973.

(42) Dasarathy, B. V. *Nearest Neighbour, NN Norm: NN Pattern Classification Techniques;* IEEE Computer Society Press: Los Alamitos, CA, 1991.

(43) Bakken, G. A., Ph.D. Thesis, Penn State University, 2001.

(44) Sutter, J. M.; Jurs, P. C. Selection of molecular structure descriptors for quantitative structure−activity relationships. In *Adaption of Simulated Annealing to Chemical Problems*; Kalivas, J. H., Ed.; Elsevier Science Publishers B. V.: Amsterdam, 1995.

(45) Sutter, J. M.; Dixon, S. L.; Jurs, P. C. Automated descriptor selection for quantitative structure−activity relationship using generalized simulated annealing. *J. Chem. Inf. Comput. Sci*. **1995**, *35*, 77−84.

(46) Wessel, M. D. Computer-Assisted Development of Quantitative Structure−Property Relationships and Design of Feature Selection Routines. Ph.D. Dissertation, Pennsylvania State University, University Park, PA, 1996.

(47) Xu, L.; Ball, J.; Dixon, S. L.; Jurs, P. C. Quantitative Structure−Activity Relationships for Toxicity of Phenols. Using Regression Analysis and Computational Neural Networks. *Environ. Toxicol. Chem.* **1994**, *13*, 841−851.

(48) Cupid, B. C.; Beddel, C. R.; Lindon, J. C.; Wilson, I. D.; Nicholson, J. K. Quantitative Structure-Metabolism Relationships for Substituted Benzoic Acids in Rabbit: Prediction of Urinary Excretion of Glycine and Glucuronide Conjugates. *Xenobiotica* **1996**, *26*, 157−176.

(49) Randic, M. On Molecular Identification Numbers. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 164−175.

(50) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.

(51) Vinogradov, S. N.; Linnell, R. H. *Hydrogen Bonding*; Van Nostrand: Reinhold: New York, 1971.

CI020045E