Classification of Multidrug-Resistance Reversal Agents Using Structure-Based Descriptors and Linear Discriminant Analysis

Gregory A. Bakken and Peter C. Jurs*

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

Received June 7, 2000

Linear discriminant analysis is used to generate models to classify multidrug-resistance reversal agents based on activity. Models are generated and evaluated using multidrug-resistance reversal activity values for 609 compounds measured using adriamycin-resistant P388 murine leukemia cells. Structure-based descriptors numerically encode molecular features which are used in model formation. Two types of models are generated: one type to classify compounds as inactive, moderately active, and active (three-class problem) and one type to classify compounds as inactive or active without considering the moderately active class (two-class problem). Two activity distributions are considered, where the separation between inactive and active compounds is different. When the separation between inactive and active classes is small, a model based on nine topological descriptors is developed that produces a classification rate of 83.1% correct for an external prediction set. Larger separation between active and inactive classes raises the prediction set classification rate to 92.0% correct using a model with six topological descriptors. Models are further validated through Monte Carlo experiments in which models are generated after class labels have been scrambled. The classification rates achieved demonstrate that the models developed could serve as a screening mechanism to identify potentially useful MDRR agents from large libraries of compounds.

Introduction

Cellular resistance to therapeutic agents is currently a major difficulty in cancer chemotherapy.[1] Often, the cellular resistance is acquired after an initial period of successful chemotherapy. Resistance of this sort is generally not attained toward a single drug but toward many structurally and functionally unrelated druglike compounds. Therefore, this phenomenon is termed multidrug resistance (MDR).[2,3]

Although the underlying mechanism is not fully understood, MDR in cells is often marked by an increase in P-glycoprotein (Pgp) on the cell surface.[3,4] Pgp acts as an energy-dependent drug efflux pump, which removes the chemotherapeutic agent from the cell and thus reduces cellular accumulation. Since the three-dimensional structure of Pgp is not yet known, it is not possible to design inhibitors to bind to the active site of the protein. Even if the three-dimensional structure of Pgp were known, it would be very difficult to design and synthesize the needed active site inhibitors. Therefore, other solutions must be sought to overcome MDR.

A promising approach is the use of multidrug-resistance reversal (MDRR) agents in conjunction with chemotherapeutic agents. Like MDR, the exact mechanism of MDRR is not known.[5] A variety of compounds have shown ability as MDRR agents,[6] including phenothiazines,[7] benzofurans,[8] quinolines,[9] sesquiterpene esters,[10] stipiamides,[11,12] and dihydrobenzopyrans and tetrahydroquinolines.[13] Many attempts have been made to elucidate structure−activity relationships (SARs) for MDRR agents,[7−10,14−17] and several such attempts have been recently summarized.[5] Any effort to determine a SAR for MDRR agents is useful in that it may lead to a viable screening procedure or to the design of new MDRR agents.

Klopman et al. developed a SAR for MDRR agents using a data set of 609 druglike compounds.[5] A subset of 351 compounds was selected and used to generate models to classify compounds as inactive or active. Biophores were identified that were important for MDRR activity by identifying structural fragments correlated with activity. Two classification models were generated, and each was validated using additional compounds. Prediction rates of 81% and 82% were obtained for the two models, demonstrating the predictive ability of the methodology used. Additional steps were taken to identify new compounds likely to be good MDRR agents. Compounds were selected by searching databases for compounds containing the biophores important in differentiating inactive and active compounds. Fourteen compounds were selected and experimentally tested to determine MDRR activity. The predicted activity for 10 of the 14 compounds matched the experimentally determined activity.

This paper presents development of classification models for MDRR activity based on structural descriptors, as opposed to structural biophores. The same set of compounds employed by Klopman et al.[5] is used here. Structure-based descriptors are used to develop classification models using linear discriminant analysis (LDA). Predictive ability of all models developed is examined using external prediction sets. Models developed could be used to screen large libraries of compounds to identify those likely to display activity as MDRR agents.

* To whom correspondence should be addressed. Phone: 814-865-3739. Fax: 814-865-3314. E-mail: pcj@psu.edu.

10.1021/jm000244u CCC: $19.00
Published on Web 10/12/2000

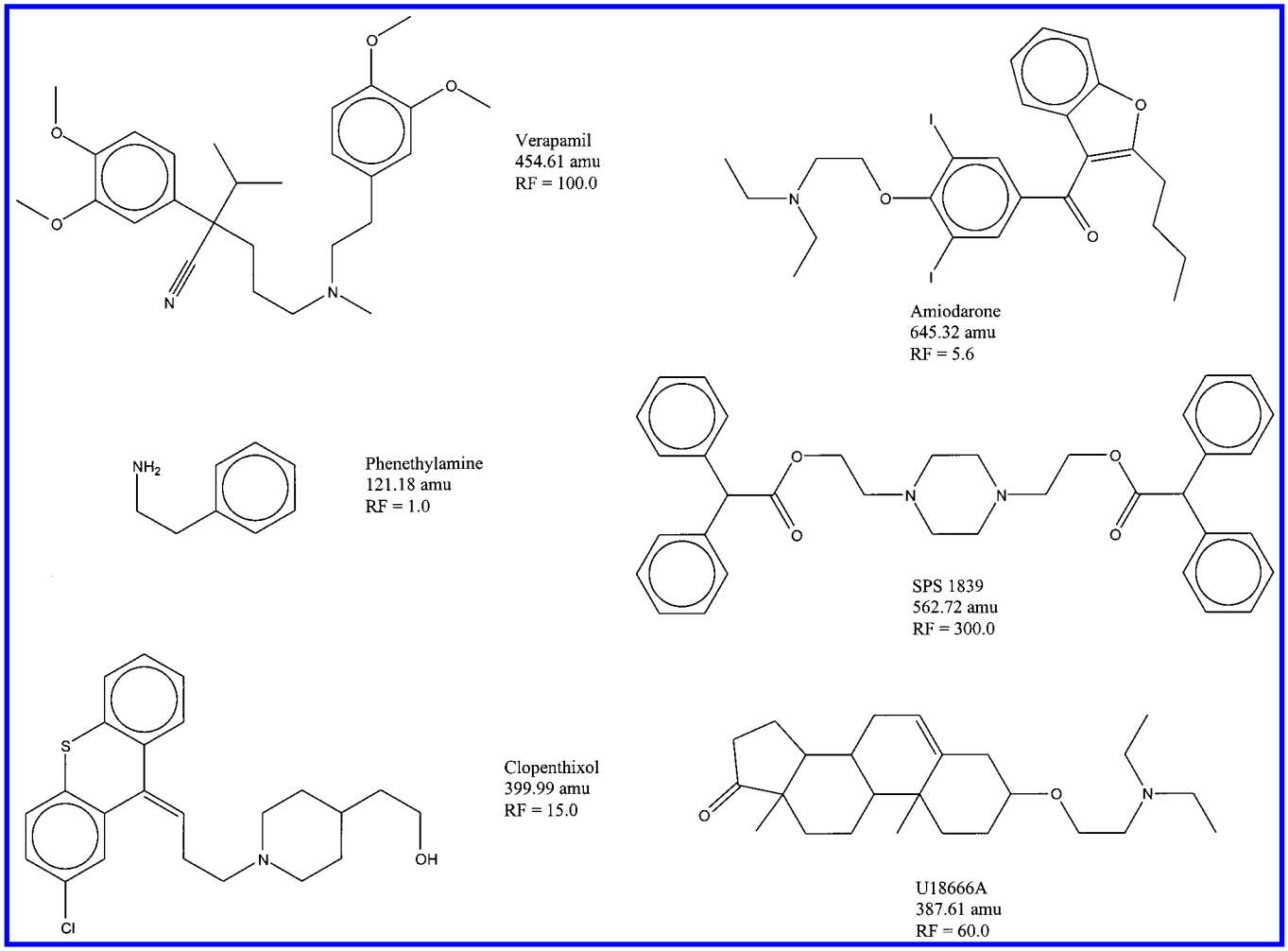

Figure 1. Example compounds demonstrating the diversity of the data set.

Experimental Section

Description of Data. To maximize the likelihood of finding a relationship between chemical structure and MDRR activity, it is beneficial to consider data collected under the same experimental conditions. Often, this results in small groups of compounds being used to develop SARs. A notable exception to this is the work by Klopman et al.[5] They developed a highly successful SAR relating the structure of 609 druglike compounds to MDRR activity using the MULTICASE (*multi*ple *c*omputer *a*utomated *s*tructure *e*valuation; MULTICASE, Beachwood, OH) program.[17,18] All experimental procedures were carried out by Ramu and Ramu, and experimental details can be found in the literature.[14−16] Briefly, ED_{50} values were collected for a line of P388 murine leukemia cells that were resistant to adriamycin (ADR). (ED_{50} values denote the drug concentration effective in inhibiting cell growth by 50%.) The ED_{50} values were collected for cells in the presence of the drug and for cells in the presence of the drug and 200 nM ADR. The drug's ability to reverse MDR (RF) was measured as: RF = (ED_{50} with no ADR)/(ED_{50} with 200 nM ADR). An RF value of 1.0 indicates no ability to reverse MDR, while large RF values indicate excellent ability to reverse MDR. The compound structures and their associated RF values were generously provided by Dr. G. Klopman.

Of the 609 compounds available, 12 structures could not be considered due to software limitations, leaving 597 compounds for analysis. For the set of 597 compounds, RF values ranged from 1.0 to 300.0. Molecular weights ranged from 121.18 (phenethylamine) to 645.32 amu (amiodarone), with an average of 368.75 amu. Figure 1 shows some example compounds demonstrating the structural diversity of the data. All 597 compounds contained nitrogen, 458 contained oxygen, 89 contained sulfur, 114 contained chlorine, 64 contained fluorine, 5 contained bromine, and 1 contained iodine. Additionally, 590 structures contained aromatic bonds, 343 contained double bonds, and 20 contained triple bonds. Structural information, in the form of SMILES notation, for all compounds is available as Supporting Information.

To generate classification models, compounds were first grouped according to MDRR activity. Klopman et al.[5] suggest labeling compounds with RF $\leq$ 2.0 as inactive, compounds with 2.0 < RF $\leq$ 4.2 as moderately active, and compounds with RF > 4.2 as active. Such a division was used in the present study and will be referred to as data set 1. When generating classification models, Klopman et al.[5] considered class labels of RF = 1.0 as inactive, 1.0 < RF < 10.0 as moderately active, and RF $\geq$ 10.0 as active. This division of compounds was also considered and will be referred to as data set 2. In general, the division of compounds into classes based on activity is somewhat arbitrary. Therefore, we considered both suggested divisions of data to demonstrate that prediction models could be generated in each case. Compounds were divided into training sets (TSETs), which were used to generate classification models, and prediction sets (PSETs), which were used to validate classification models. Table 1 shows the distribution of compounds into classes for data sets 1 and 2. For both data sets, classification models were generated for the two-class problem (inactive/active) and the three-class problem (inactive/moderate/active).

This study was carried out using the automated data analysis and pattern recognition toolkit (ADAPT) software system.[19,20] The genetic algorithm (GA)[21,22] feature selection routine and the LDA[23,24] routine were written in-house. All computations were performed on a DEC 3000 AXP model 500 workstation. Methods used to develop classification models can be broken into four basic steps: (1) structure entry and

Table 1. Distribution of Compounds into TSETs and PSETs for Data Sets 1 and 2

	TSET	PSET
Data Set 1		
inactive (RF ≤ 2.0)	146	77
active (RF > 4.2)	199	100
moderate (2.0 < RF ≤ 4.2)	52	23
Data Set 2		
inactive (RF = 1.0)	99	48
active (RF ≥ 10.0)	130	64
moderate (1.0 < RF < 10.0)	168	88

modeling, (2) descriptor generation, (3) objective feature selection, and (4) model formation and validation.

(1) Structure Entry and Modeling. Initial three-dimensional coordinates for non-hydrogen atoms of the compounds were obtained using ChemDraw (CambridgeSoft Corp., Cambridge, MA). The compounds were then examined in HyperChem (Hypercube, Inc., Waterloo, Ontario), and hydrogens were added and coordinates were refined. Compound structures were stored as connection tables which contained information about atom types, bond angles, and bond multiplicity. To obtain more accurate three-dimensional geometries, which are necessary for some descriptor calculations, the refined coordinates were passed to the semiempirical molecular orbital program MOPAC.[25] A PM3 Hamiltonian[26] was selected for geometry optimization.

(2) Descriptor Generation. To relate molecular structure to MDRR activity, descriptors that accurately encode the structural features responsible for the observed activity are necessary. Therefore, it is important that an information-rich pool of descriptors be available. To generate such a pool, 220 structure-based descriptors were calculated for all 597 compounds. The calculated descriptors can be labeled as topological, geometric, electronic, and polar surface descriptors. Of the 220 descriptors calculated in this study, 138 were topological, 24 were geometric, 10 were electronic, and 48 were polar surface descriptors.

Topological descriptors[27−32] are calculated based on a two-dimensional sketch of the compound. A significant advantage of this descriptor type is that geometry optimization of the structures is not required. This is especially important for large, druglike molecules for which geometry optimization can be a very time-consuming process and accurate geometries are often difficult to obtain. Topological descriptors calculated included molecular connectivity indices, molecular distance edge descriptors, κ indices, and fragment descriptors. Molecular connectivity indices and molecular distance edge descriptors encoded information about molecular size and degree of branching. κ indices provided information about molecular shape using only the two-dimensional compound sketch. Fragment descriptors simply represented counts of atoms, atom types, bonds, rings, etc.

Geometric descriptors[33,34] are more computationally intensive because they require accurate three-dimensional coordinates for the compounds. Examples of geometric descriptors included moments of inertia, gravitational index, and solvent-accessible surface areas and volumes. The gravitational index encoded information about compound size. Surface areas and volumes provided information about the ability of the compound to interact with solvent, e.g., water.

Like geometric descriptors, electronic descriptors[26] are frequently very computationally demanding. These descriptors encoded the electronic environment of the compounds. Calculated descriptors included the electric dipole moment, energy of the highest occupied molecular orbital, and energy of the lowest unoccupied molecular orbital.

Polar surface descriptors, also called charged partial surface area descriptors, represent a hybrid or combination of the three previous descriptor types. These descriptors combine information from at least two of the classes previously described. For example, partial atomic charges (electronic information) were combined with solvent-accessible surface area (geometric information) to form charged partial surface area descriptors.[35] Additional polar surface descriptors encoded hydrogen-bonding information for the compounds. Polar surface information may help to measure the hydrophobic/hydrophilic and/or lipophobic/lipophilic nature of compounds.

(3) Objective Feature Selection. The process of feature selection entails pruning the descriptor pool through objective and subjective means. Objective feature selection involves eliminating descriptors based solely on descriptor values. Subjective feature selection, discussed in the next section, utilizes dependent variable information (class label). To avoid chance correlations, objective means should be used to reduce the descriptor pool to a reasonable level before using subjective means. In practice, the ratio of descriptors to TSET observations should be equal to or less than 0.6 before subjective feature selection.

Objective feature selection was carried out using the 397 TSET observations. Any descriptor containing identical values for 90% or more of the TSET observations was eliminated. Such descriptors provided no information useful for differentiating compound classes. Additionally, pairwise correlations were calculated for all descriptors. One of any two descriptors with a correlation above 0.9 was eliminated. These two measures reduced the initial pool of 220 descriptors to 100 descriptors. A reduced pool of 100 descriptors was deemed acceptable since the ratio of descriptors to TSET observations was well below 0.6 for all cases investigated. Of the 100 descriptors in the reduced pool, 60 were topological in nature. The pool of 60 topological descriptors alone was also used to generate models. As mentioned previously, models based on only topological descriptors offer the advantage of not requiring geometry-optimized structures.

(4) Model Formation and Validation. The two reduced descriptor pools were screened using genetic algorithm[21,22] evolutionary optimization. Descriptor subsets were examined to see if they could develop classifiers to determine MDRR activity level. Models were formed using LDA,[23,24] *k*-nearest neighbor analysis (*k*NN), and radial basis function neural networks (RBFNN).[36] For the work described here, LDA provided better results than *k*NN and RBFNN. Therefore, LDA will be the only technique discussed, and results obtained using LDA will be presented.

LDA is a supervised classification technique that maximizes separation between class means in some descriptor space, relative to standard deviation. Discriminants were generated using TSET compounds. Model sizes ranging from 5 to 15 descriptors were investigated. For each model size, a GA was used to search the descriptor space for the subset of descriptors that produced the lowest COST function (see below for discussion of COST). Models of various size were compared to find the model providing the lowest COST with the fewest descriptors. Once selected, the optimal model was used to classify compounds in the PSET to verify the ability of the model to generalize.

The COST function computed for each descriptor subset was designed to meet two goals: (1) ensure that the model would generalize well and (2) penalize models that provided good prediction accuracy for one class at the cost of another class. To accomplish the first goal, a leave-10%-out cross-validation approach was used. Each time a descriptor subset was examined, 10% of the TSET compounds were randomly selected to serve as a cross-validation set (CVSET) to validate discriminants and the remaining 90% of the compounds were used to generate discriminants. For each descriptor subset, this process was repeated 10 times, such that each compound in the TSET appeared in the CVSET exactly one time. For each descriptor subset, the COST was computed as the average over the 10 TSET/CVSET combinations. For each TSET/CVSET combination, COST was computed as $TSET_w + 0.5|TSET_w - CVSET_w|$, where $TSET_w$ denoted the percent incorrect classification *for the most poorly predicted class* in the TSET and $CVSET_w$ denoted the same thing for the CVSET compounds. COST was computed with $TSET_w$ and $CVSET_w$ to work toward goal 2 above.

This model selection approach produced robust models. By using a leave-*N*%-out approach with the GA, the ability of each model to generalize to compounds not used in discriminant generation was tested during the descriptor selection stage. This helped prevent selection of models only able to classify TSET compounds. Additionally, models were avoided which reduced incorrect classification overall by increasing misclassification for a particular class. This may be a concern when there is an uneven distribution in the number of compounds in each class. Such uneven distributions were present in all analyses presented here.

Results and Discussion

Subsets of 5−15 descriptors were examined. The smallest descriptor subset that produced an acceptably low COST value was selected as optimal. The descriptor values for the TSET compounds were then used to generate discriminants. Once generated, the discriminants were used to classify all compounds in the TSET and the percent correct classification was calculated. The compounds in the PSET were then classified, and the percent correct classification for these compounds served as validation of the model.

Data Set 1: Two-Class Problem. The first problem investigated for data set 1 was the generation of models to differentiate inactive and active compounds. Models were formed from the reduced pool of 100 descriptors as well as the subset of 60 topological descriptors. The optimal model formed from the reduced pool of 100 descriptors was a seven-descriptor model. Of the seven descriptors, four were topological, one was geometric, and two were polar surface descriptors. Of the 345 TSET compounds, 82.0% were correctly classified. The correct classification rate for the 177 PSET compounds was 81.9%, which clearly demonstrates that the model generated is capable of classifying compounds not used in model formation.

A second model was generated to address the inactive/active problem. This model was formed by only selecting from the 60 topological descriptors in the reduced pool. Models based solely on topological descriptors offer the distinct advantage of not requiring geometry optimization of the structures. Using only topological descriptors, a nine-descriptor model was selected as optimal. With this model, 82.0% of the TSET compounds were correctly classified. For the PSET compounds, the classification rate rose to 83.1%.

The nine topological descriptors selected are shown in Table 2. The three descriptors NO3, NN4, and NAB15 represent counts of oxygen atoms, nitrogen atoms, and aromatic bonds, respectively. The appearance of these descriptors was not surprising based on the characterization of the data set provided previously. KAPA6 denotes the path 3 κ index (corrected for atom type), which compares a compound containing *N* atoms to the most and least branched forms of an *N* atom compound.[27] This descriptor provides information about molecular shape based on a two-dimensional compound sketch. The two descriptors ALLP4 and S4C9 encode information about topological complexity and branching of the compound structures. ALLP4 is the total weighted number of paths relative to the total number of atoms,[32] and S4C9 denotes a simple 4th order cluster term.[37] The molecular distance edge terms, MDE13 and MDE44, describe connection information between primary and tertiary carbons and pairs of quaternary carbons, respectively.[28] ESUM2 denotes the sum of electrotopological state values over all heteroatoms.[38] Electrotopological state values provide information about intermolecular interactions.

Table 2. Nine Topological Descriptors Defining the Optimal Two-Class Model for Data Set 1

descriptor[a]	discrim coeff	range		average		rel var[b]
		inactive	active	inactive	active	
NO3	0.186	0−8	0−10	2.12	2.20	1.56
NN4	0.111	1−8	1−9	2.49	2.58	1.22
NAB15	−0.011	0−18	0−24	10.24	13.08	1.21
KAPA6	−0.333	1.27−11.02	2.53−13.25	4.77	6.08	0.72
ALLP4	−3.766	1.71−2.52	1.79−2.47	2.05	2.12	0.01
S4C9	0.198	0−0.40	0−0.58	0.02	0.04	0.20
MDE13	−0.039	0−34.38	0−28.09	3.71	6.19	5.43
MDE44	0.022	0−52.48	0.20−27.57	6.66	8.40	5.97
ESUM2	−0.008	2.16−76.62	2.24−96.10	29.75	35.84	10.71

[a] Explanation: NO3, number of oxygen atoms in the compound; NN4, number of nitrogen atoms in the compound; NAB15, number of aromatic bonds in the compound; KAPA6,[27] κ index based on path lengths of 3 with correction (α) for atom type, [(number of atoms + α − 1)(number of atoms + α − 2)^2] ÷ [(number of 3-bond fragments + α)^2]; ALLP4,[32] total weighted number of paths in the compound ÷ total number of atoms in the compound; S4C9,[37] simple 4th order cluster molecular connectivity; MDE13,[28] molecular distance edge term between primary and tertiary carbons; MDE44,[28] molecular distance edge term between quaternary carbons; ESUM2,[38] summation of electrotopological state indices over all heteroatoms in a compound. [b] Relative variance is the variance divided by the mean for a descriptor using all compounds from both classes.

Table 2 also shows the range and average for the inactive and active classes for each of the nine descriptors, as well as the relative variance for each descriptor. Interestingly, the average descriptor value for active compounds is always larger than the average for inactive compounds, although in several cases the difference is insignificant. For NO3, NN4, ALLP4, and S4C9, descriptor ranges and averages are nearly equivalent for inactive and active compounds. Descriptors ALLP4 and S4C9 are somewhat suspect due to the low relative variance for each. Removing these two descriptors reduced classification rates to 73.9% and 79.1% for the TSET and PSET, respectively, which indicates that the two descriptors are providing useful information.

To ensure that the results above were not due to chance, Monte Carlo experiments were conducted in which models were generated after scrambling of class labels. The class labels were randomly scrambled for the TSET and PSET compounds 10 times. Each time, a GA was used to select the optimal nine-topological descriptor model. Discriminants were calculated, and the TSET and PSET compounds, with scrambled class labels, were classified. For the TSET, the 10 models produced a classification accuracy of 59.9 ± 2.3%. Classification of the PSET compounds produced a rate of 50.6 ± 3.6%. Based on the distribution of compounds (see Table 1), random assignment would produce classification accuracy of 51.2%, very close to the average of the 10 Monte Carlo experiments. This result clearly demonstrates that the predictive ability of the original nine-topological descriptor model is not due to chance.

Data Set 1: Three-Class Problem. Models were also generated for data set 1 to classify samples as inactive, moderately active, or active using the distribution shown in Table 1. When all 100 descriptors in the reduced pool were considered, a six-descriptor model

Table 3. Confusion Matrix for the TSET and PSET Compounds Using the Optimal Eight-Topological Descriptor Model for the Three-Class Problem for Data Set 1

a. TSET Compounds

	predicted class			
actual class	inactive	moderate	active	% correct
inactive	107	20	19	73.3
moderate	7	36	9	69.2
active	25	40	134	67.3

b. PSET Compounds

	predicted class			
actual class	inactive	moderate	active	% correct
inactive	55	15	7	71.4
moderate	6	12	5	52.2
active	8	21	71	71.0

was selected as optimal. Interestingly, five of the six descriptors selected were topological and one was electronic. The correct classification rate was 67.8% for the TSET compounds and 67.0% for the PSET compounds. The classification rates are significantly lower than those obtained for the inactive/active situations, but the problem is a more difficult one. Additionally, the rates are higher than the expected random result of 40.4%.

Classification models generated from only topological descriptors provided similar prediction accuracy. The optimal model, containing eight topological descriptors, produced a 69.8% correct classification rate for the TSET compounds. Table 3a shows the confusion matrix for the TSET compounds using this model. Also shown are the correct classification rates for each of the three classes. This demonstrates that the COST function employed found a descriptor subset not biased for or against any one class for the TSET compounds. Because the moderately active class is so small, many prediction models might misclassify the majority of the moderately active compounds to increase accuracy for the inactive and active classes, thereby increasing overall classification accuracy.

For the PSET compounds, overall classification accuracy (69.0%) was close to that obtained for the TSET compounds, but the accuracy for individual classes was quite different. Table 3b shows that classification accuracy is good for the active and inactive classes but extremely poor for the moderately active class. This indicates that, although the COST function was able to find a descriptor subset that performed equally well for all three classes for the TSET, the model did not generalize well in terms of accuracy for individual classes. This validation failure for the moderately active class was characteristic of all three-class models developed for data set 1. This may be an unavoidable consequence of having highly uneven class populations, or it may be that the COST function employed was not adequate for this situation.

Monte Carlo experiments were conducted using scrambled class labels. The class labels were randomly scrambled 10 times each for the TSET and PSET compounds, and the optimal eight-topological descriptor model was selected each time. Classification accuracy was 44.5 ± 4.8% for the TSET compounds and 35.3 ± 4.6% for the PSET compounds. The classification rate for the PSET compounds is not within one standard deviation of the expected random result of 40.4%. This phenomenon was reproducible in multiple Monte Carlo experiments. Further investigation revealed that the combination of the uneven class distribution and the COST function produced this result. Additional Monte Carlo experiments were conducted with a COST function equal to the percent incorrect over all TSET observations. When the 10 Monte Carlo experiments were repeated with this COST function, the TSET and PSET errors were 50.7 ± 3.0% and 39.9 ± 4.7%, respectively. The PSET classification rate is now very close to the expected random result.

Data Set 2: Two-Class Problem. Data set 2 (see Table 1) was chosen based on guidelines Klopman et al. used in their classification study.[5] Inactive compounds are those that produced no change in ED_{50}, and active compounds are those that decrease ED_{50} by an order of magnitude or more. This produces greater separation between active and inactive compounds for the two-class problem and also increases the size of the moderately active class for the three-class problem.

For the two-class problem, a seven-descriptor model was selected as optimal when all descriptor types were available in the reduced pool. Four of the descriptors were topological, one was electronic, and two were polar surface descriptors. The classification rates were 90.0% and 91.1% for the TSETs and PSETs, respectively. This result compares very favorably with previous results obtained using very different methodology.[5] For the 130 active TSET compounds, the average RF value was 30.3 and values ranged from 10.0 to 300.0. Thirteen of the 130 active compounds were labeled as inactive. For these 13 compounds, the average RF was 12.7, with a range of 10.0−20.0. This demonstrates that the mislabeling of active compounds as inactive occurs mainly for compounds close to the cutoff for activity. Of the 64 active PSET compounds, five were labeled inactive. RF values for these ranged from 10.0 to 50.0, with an average of 18.0. Bevanitolol (Figure 2a) was the compound with RF = 50.0 that was labeled as inactive. Other compounds similar in structure to bevanitolol were present, and the descriptor values for bevanitolol were not outside the range of the TSET compounds. There was no clear reason bevanitolol was not correctly classified.

A classification model was also generated using only topological descriptors. The six-descriptor model selected as optimal is described in Table 4. Using this model, 90.8% of the TSET compounds were correctly classified. Additionally, 92.0% of the PSET compounds were correctly classified. This model is very encouraging in that all descriptors used are topological and classification rates are greater than 90%. Such a model provides a viable means for screening large libraries of compounds.

Three of the six descriptors chosen, NC2, NO3, and NS5, represented counts of carbon atoms, oxygen atoms, and sulfur atoms, respectively. Similar descriptor types were also selected for the data set 1 model shown in Table 2. ALLP2 is the total number of paths divided by the total number of atoms,[32] which simply encodes the number of non-hydrogen atoms. N6P17 is the number of 6th order paths,[37] which provides information about topological complexity and structure branching. SYMM15 denotes a descriptor designed to encode information

Table 4. Six Topological Descriptors Defining the Optimal Two-Class Model for Data Set 2

		range		average		
descriptor[a]	discrim coeff	inactive	active	inactive	active	rel var[b]
NC2	−0.143	9−30	17−38	17.01	24.98	1.54
NO3	0.100	0−6	0−10	2.19	2.55	1.31
NS5	−0.270	0−1	0−2	0.07	0.18	1.14
ALLP2	0.012	9.00−204.37	22.80−163.06	41.22	68.33	21.84
N6P17	−0.010	8−184	43−187	66.68	103.72	19.15
SYMM15	−1.361	0.35−1.00	0.29−1.00	0.77	0.77	0.04

[a] Explanation: NC2, number of carbon atoms in the compound; NO3, number of oxygen atoms in the compound; NS5, number of sulfur atoms in the compound; ALLP2,[32] total number of paths in the compound ÷ total number of atoms in the compound; N6P17,[37] number of 6th order paths; SYMM15,[39] molecular symmetry descriptor calculated using chemical environment defined by atoms up to 5 bonds away. [b] Relative variance is the variance divided by the mean for a descriptor using all compounds from both classes.

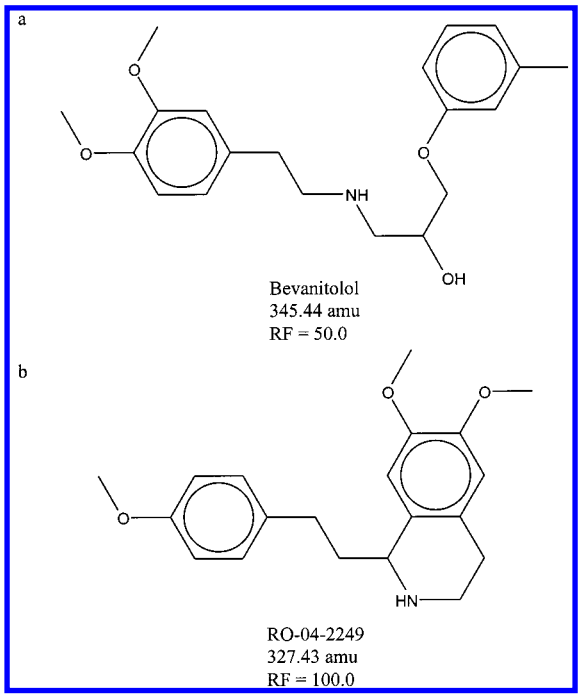

Figure 2. Active compounds from data set 2 incorrectly labeled as inactive. (a) Bevanitolol was a PSET compound labeled as inactive using both the seven-descriptor model and the six-topological descriptor model. For bevanitolol, the molecular weight was 345.44 amu and the RF was 50.0. (b) RO-04-2249 was a TSET compound incorrectly labeled as inactive using the six-topological descriptor model. For RO-04-2249, the molecular weight was 327.43 amu and the RF was 100.0.

about molecular symmetry based on topology.[39] The number of unique atoms is divided by the total number of atoms, where a unique atom is defined based on chemical environment up to five bonds away from the atom of interest.

Table 4 shows the average, range, and relative variance data for the six descriptors. Here again, average descriptor values tend to be higher for active compounds. SYMM15 is an exception, since the average value is the same for inactive and active compounds. Additionally, the relative variance for this descriptor is quite low, calling into question whether discriminating information is present. Removing SYMM15 from the model reduced the TSET classification rate to 85.2% and the PSET rate to 83.9%. Therefore, SYMM15 must contain some information useful in classifying inactive and active compounds.

With this model, 12 of the 130 active TSET compounds were labeled as inactive. For these 12, the RF values ranged form 10.0 to 100.0 with an average value of 21.0. Other than the one compound with RF = 100.0, no mislabeled compound had RF > 22.2. The compound with RF = 100.0 is shown in Figure 2b. As with bevanitolol, no reason for the misclassification was apparent. Four of the 64 active PSET compounds were also labeled as inactive. The RF values for these four compounds ranged from 10.0 to 50.0, with an average of 21.3. Again, the only PSET compound with a relatively high RF value labeled as inactive was bevanitolol.

Monte Carlo experiments were conducted to further validate the models developed for data set 2. The class labels were randomly scrambled for the TSET and PSET compounds 10 times, and a GA was used to select the optimal six-topological descriptor model each time. The 10 models produced a classification accuracy of 60.3 ± 2.7% for the TSET compounds and 47.1 ± 5.1% for the PSET compounds. In this case, random assignment would produce a 50.9% correct classification rate. This clearly demonstrates that the models developed are not identifying chance relationships.

Data Set 2: Three-Class Problem. As with data set 1, models were generated to classify compounds as inactive, moderately active, and active. The distribution of compounds into these classes in shown in Table 1. Compared to data set 1, the three classes are more evenly represented. Using the reduced pool of 100 descriptors, an eight-descriptor model was selected as optimal. The correct classification rates for the TSET and PSET were 72.3% and 61%, respectively. The low PSET classification rate was somewhat disappointing. Examination of confusion matrices (not shown) demonstrated that the main source of confusion was between moderately active compounds and active compounds. For the TSET, 54.5% of the errors were due to confusion between moderately active and active compounds, as were 60.5% of the errors for the PSET.

When models were formed using only topological descriptors, a six-descriptor model was selected as optimal. The TSET classification rate was 67.8% and the PSET rate was 62.5%. Examination of the confusion matrices in Table 5 illustrates that misclassification of moderately active compounds is responsible for a large part of the error. Additionally, classification of active compounds as moderately active occurs frequently. Table 5 shows that the inactive compounds are well-predicted, while both the moderately active and active compounds are poorly predicted. This result is different than that seen in Table 3 for data set 1. Although the classification rates for the three-class problem are somewhat low, they are still significantly better than

Table 5. Confusion Matrix for the TSET and PSET Compounds Using the Optimal Six-Topological Descriptor Model for the Three-Class Problem for Data Set 2

a. TSET Compounds

	predicted class			
actual class	inactive	moderate	active	% correct
inactive	78	14	7	78.8
moderate	32	107	29	63.7
active	9	37	84	64.6

b. PSET Compounds

	predicted class			
actual class	inactive	moderate	active	% correct
inactive	39	8	1	81.3
moderate	20	47	21	53.4
active	2	23	39	60.9

the expected random results of 34.9%. Additionally, Monte Carlo experiments with scrambled class labels and topological descriptors produced classification rates of $41.2 \pm 2.1\%$ and $33.3 \pm 4.7\%$ for the TSET and PSET, respectively, providing further validation of the predictive ability of the six-topological descriptor model.

Conclusions

Linear discriminant analysis was used to generate classification models based on topological descriptors only and models based on topological, geometric, electronic, and/or polar surface descriptors. In general, models employing only topological descriptors provided predictive accuracy as good as or better than models developed with multiple descriptor types. This may be a result of the geometries used for the compounds. It is difficult to get accurate geometries for large molecules. Several low-energy conformation are available, and it is not possible to identify which conformation binds to Pgp. Additionally, the calculated geometries were for compounds in vacuo. Conformations are probably quite different when the compounds interact with Pgp.

Although somewhat surprising, the success of models based on topological descriptors is encouraging. Computational requirements for topological descriptors are generally low and therefore amendable to screening large libraries of complex compounds. For data set 2, a six-topological descriptor model was generated that correctly classified 103 of 112 PSET compounds. Additionally, any misclassification of active compounds as inactive occurred mainly for compounds with RF values close to the cutoff for active compounds. Therefore, this model could serve as a screening mechanism to rapidly select compounds from a large library for further consideration as MDRR agents.

Klopman et al.[5] developed classification models capable of correctly classifying 82% of external PSET compounds for data set 2 using the MULTICASE program. These models were generated based on structural biophores identified as important for activity. The identified biophores provided useful information for generating additional MDRR agents. The methodology presented here provided a higher classification rate (92.0%) for external PSET compounds, at the cost of model interpretation. The relationship between structural descriptors and MDRR activity is not easily defined. However, each method provides an easily implemented and computationally efficient means of selecting MDRR candidates based on molecular structure.

The high prediction accuracy of the models developed indicates that structure-based descriptors do encode information useful in differentiating MDRR agents based on activity levels. In an attempt to increase predictive accuracy, the partition coefficient (log *P*) was added to the pool of topological descriptors, and new models were formed for data set 2. Although log *P* was selected by the GA procedure for inclusion in some model sizes, the COST value for these models remained comparable to the COST values obtained when log *P* was not included. Based on the similarity of the COST values, it was not surprising to find that classification accuracy for models with log *P* included was no better than accuracy of models without log *P*. This finding suggests that, for the present problem, any information useful in discriminating active and inactive compounds provided by log *P* is already encoded in other descriptors present in the reduced descriptor pool.

Acknowledgment. The authors are grateful to Dr. G. Klopman for providing compound structures and RF values.

Supporting Information Available: Compound structures, in SMILES notation, along with RF values. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) *Multidrug Resistance in Cancer Cells: Molecular, Biochemical, Physiological and Biological Aspects*; Gupta, S., Tsuruo, T., Eds.; John Wiley & Sons: Chichester, 1996.
(2) Biedler, J. L.; Riehm, H. Cellular Resistance to Actinomycin D in Chinese Hamster Cells in Vitro: Cross-Resistance, Radioautographic, and Cytogenetic Studies. *Cancer Res.* **1970**, *30*, 1174−1184.
(3) Endicott, J. A.; Ling, V. The Biochemistry of P-Glycoprotein-mediated Multidrug Resistance. *Annu. Rev. Biochem.* **1989**, *58*, 137−171.
(4) Gottesman, M. M.; Pastan, I. Biochemistry of Multidrug Resistance Mediated by the Multidrug Transporter. *Annu. Rev. Biochem.* **1993**, *62*, 385−427.
(5) Klopman, G.; Shi, L. M.; Ramu, A. Quantitative Structure−Activity Relationship of Multidrug Resistance Reversal Agents. *Mol. Pharmacol.* **1997**, *52*, 323−334.
(6) Berger, D.; Citarella, R.; Dutia, M.; Greenberger, L.; Hallett, W.; Paul, R.; Powell, D. Novel multidrug resistance reversal agents. *J. Med. Chem.* **1999**, *42*, 2145−2161.
(7) Pajeva, I.; Wiese, M. Molecular modeling of phenothiazines and related drugs as multidrug resistance modifiers: A comparative molecular field analysis study. *J. Med. Chem.* **1998**, *41*, 1815−1826.
(8) Ecker, G.; Chiba, P.; Hitzler, M.; Schmid, D.; Visser, K.; Cordes, H. P.; Csollei, J.; Seydel, J. K.; Schaper, K. J. Structure−activity relationship studies on benzofuran analogues of propafenone-type modulators of tumor cell multidrug resistance. *J. Med. Chem.* **1996**, *39*, 4767−4774.
(9) Suzuki, T.; Fukazawa, N.; Sannohe, K.; Sato, W.; Yano, O.; Tsuruo, T. Structure−activity relationship of newly synthesized quinoline derivatives for reversal of multidrug resistance in cancer. *J. Med. Chem.* **1997**, *40*, 2047−2052.
(10) Kim, S. E.; Kim, H. S.; Hong, Y. S.; Kim, Y. C.; Lee, J. J. Sesquiterpene esters from *Celastrus orbiculatus* and their structure−activity relationship on the modulation of multidrug resistance. *J. Nat. Prod.* **1999**, *62*, 697−700.
(11) Andrus, M. B.; Lepore, S. D. Synthesis of stipiamide and a new multidrug resistance reversal agent, 6,7-dehydrostipiamide. *J. Am. Chem. Soc.* **1997**, *119*, 2327−2328.
(12) Andrus, M. B.; Turner, T. M.; Asgari, D.; Li, W. K. The synthesis and evaluation of a solution-phase indexed combinatorial library of nonnatural polyenes for multidrug resistance reversal. *J. Org. Chem.* **1999**, *64*, 2978−2979.
(13) Hiessbock, R.; Wolf, C.; Richter, E.; Hitzler, M.; Chiba, P.; Kratzel, M.; Ecker, G. Synthesis and in vitro multidrug resistance modulating activity of a series of dihydrobenzopyrans and tetrahydroquinolines. *J. Med. Chem.* **1999**, *42*, 1921−1926.

(14) Ramu, N.; Ramu, A. Circumvention of Adriamycin Resistance by Dipyridamole Analogues: A Structure−Activity Relationship Study. *Int. J. Cancer* **1989**, *43*, 487−491.
(15) Ramu, A.; Ramu, N. Reversal of multidrug resistance by phenothiazines and structurally related compounds. *Cancer Chemother. Pharmacol.* **1992**, *30*, 165−173.
(16) Ramu, A.; Ramu, N. Reversal of multidrug resistance by bis-(phenylalkyl)amines and structurally related compounds. *Cancer Chemother. Pharmacol.* **1994**, *34*, 423−430.
(17) Klopman, G.; Srivastava, S.; Kolossvary, I.; Epand, R. F.; Ahmed, N.; Epand, R. M. Structure−Activity Study and Design of Multidrug-resistant Reversal Compounds by a Computer Automated Structure Evaluation Methodology. *Cancer Res.* **1992**, *52*, 4121−4129.
(18) Klopman, G. MULTICASE. A hierarchical computer automated structure evaluation program. *Quant. Struct.-Act. Relat.* **1992**, *11*, 176−184.
(19) Jurs, P. C.; Chow, J. T.; Yuan, M. In *Computer-Assisted Drug Design*; Olson, E. C., Christofferson, R. E., Eds.; The American Chemical Society: Washington, DC, 1979.
(20) Stuper, A. J.; Brugger, W. E.; Jurs, P. C. *Computer-Assisted Studies of Chemical Structure and Biological Function*; Wiley: New York, 1979.
(21) 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.
(22) Kimura, T.; Hasegawa, K.; Funatsu, 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.
(23) Huberty, C. J. *Applied Discriminant Analysis*; John Wiley & Sons: New York, 1994.
(24) Johnson, R. A.; Wichern, D. W. *Applied Multivariate Statistical Analysis*; Prentice-Hall: Englewood Cliffs, NJ, 1982.
(25) Stewart, J. J. P. *Mopac 6.0, Quantum Chemistry Program Exchange*; Indiana University.
(26) Stewart, J. J. P. MOPAC: A Semiempirical Molecular Orbital Program. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 1−105.
(27) Kier, L. B. Shape Indexes of Orders One and Three from Molecular Graphs. *Quant. Strucut.-Act. Relat.* **1986**, *5*, 1−7.
(28) Liu, S.; Cao, C.; Li, Z. Approach to Estimation and Prediction for Normal Boiling Point (NBP) of Alkanes Based on a Novel Molecular Distance-Edge (MDE) Vector, *λ*. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 387−394.
(29) 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.
(30) Randic, M.; Brissey, G. M.; Spencer, R. B.; Wilkins, C. L. Search for all self-avoiding paths for molecular graphs. *Comput. Chem.* **1979**, *3*, 5−13.
(31) Randic, M. On Molecular Identification Numbers. *J. Chem. Inf. Comput. Sci.* **1984**, *4*, 162−175.
(32) Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69*, 17−20.
(33) Bondi, A. van der Waals Volumes and Radii. *J. Phys. Chem.* **1964**, *68*, 441−451.
(34) Stouch, T. R.; Jurs, P. C. A Simple Method for the Representation, Quantification, and Comparison of the Volumes and Shapes of Chemical Compounds. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 4−12.
(35) Stanton, D. T.; Jurs, P. C. Development and Use of Charged Partial Surface Area Structural Descriptor in Computer-Assisted Quantitative Structure−Property Relationship Studies. *Anal. Chem.* **1990**, *62*, 2323−2329.
(36) Wan, C.; Harrington, P. B. Self-Configuring Radial Basis Function Neural Networks for Chemical Pattern Recognition. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1049−1056.
(37) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; Research Studies Press Ltd.: Hertfordshire, England, 1986.
(38) Kier, L. B.; Hall, L. H. The E-State as an Extended Free Valence. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 548−552.
(39) Small, G. W. Ph.D. The Pennsylvania State University, 1984.

JM000244U