Spline-Fitting with a Genetic Algorithm: A Method for Developing Classification Structure−Activity Relationships

Jeffrey J. Sutherland,[†] Lee A. O'Brien,[‡] and Donald F. Weaver*[,§]

Departments of Chemistry and Pathology, Queen's University, Kingston, Ontario, Canada K7L 3N6, and Departments of Medicine (Neurology) and Chemistry and School of Biomedical Engineering, Dalhousie University, Halifax, Nova Scotia, Canada B3H 4J3

Received July 14, 2003

Classification methods allow for the development of structure−activity relationship models when the target property is categorical rather than continuous. We describe a classification method which fits descriptor splines to activities, with descriptors selected using a genetic algorithm. This method, which we identify as SFGA, is compared to the well-established techniques of recursive partitioning (RP) and soft independent modeling by class analogy (SIMCA) using five series of compounds: cyclooxygenase-2 (COX-2) inhibitors, benzodiazepine receptor (BZR) ligands, estrogen receptor (ER) ligands, dihydrofolate reductase (DHFR) inhibitors, and monoamine oxidase (MAO) inhibitors. Only 1-D and 2-D descriptors were used. Approximately 40% of compounds in each series were assigned to a test set, "cherry-picked" from the complete set such that they lie outside the training set as much as possible. SFGA produced models that were more predictive for all but the DHFR set, for which SIMCA was most predictive. RP gave the least predictive models for all but the MAO set. A similar trend was observed when using training and test sets to which compounds were randomly assigned and when gradually eliminating compounds from the (designed) training set. The stability of models was examined for the random and reduced sets, where stability means that classification statistics and the selected descriptors are similar for models derived from different sets. Here, SIMCA produced the most stable models, followed by SFGA and RP. We show that a consensus approach that combines all three methods outperforms the single best model for all data sets.

INTRODUCTION

Quantitative structure−activity relationships (QSAR) attempt to correlate the biological activities of compounds with their structural attributes, to help elucidate the mechanism by which they act and to predict the activities of novel derivatives. Traditionally, QSAR has been applied to a relatively small number of *active* congeneric compounds (<100), with activities varying continuously. These studies have often been retrospective or used to make predictions on similar derivatives. With the advent of combinatorial chemistry and high-throughput screening (HTS), the focus of QSAR has shifted toward virtual screening of compound libraries at the early stages of the drug discovery process.[1] This has been coupled to the increasing use of diversity-based approaches for assembling the libraries.[2,3] HTS data sets are larger and contain more diverse structures than those used in conventional QSAR, with biological activities often expressed in categorical form (e.g. active or inactive). This has stimulated the development of classification methods capable of dealing effectively with such data sets.

In addition, classification methods can be useful for studying more traditional QSAR sets for which activities are categorical in nature (e.g. taste, toxicity) or when they are compiled from a variety of sources, with different methodologies employed to obtain the activities. In QSAR studies, compounds with activities expressed in indeterminate form (e.g. $IC_{50} > 1\ \mu M$) are often discarded. These can be included in the development of classification models that benefit from the increased extent of chemical space covered by the training set. For compounds predicted to be active by the classification model, the prediction can be refined with a conventional QSAR model.[4,5]

Most classification methods are based on clustering, partitioning, or statistical approaches.[6] Statistical-based methods, by which a model of activity is developed, include linear discriminant analysis[7] (LDA), soft independent modeling by class analogy[8,9] (SIMCA), recursive partitioning[10,11] (RP), recursive mean partitioning,[12] and binary QSAR.[13,14] LDA, SIMCA, and RP are well-established methods that have been applied to a large number of SAR studies. In a comparative classification study, LDA and RP were found to perform better than hierarchical clustering.[15]

Nonlinear effects are more prominent in classification sets than in QSAR sets containing only active compounds. For example, activity may be linearly related to the volume of a compound within the chemical space of active compounds. When inactive compounds are included, the relation may be subject to threshold effects, as compounds with volumes above or below certain values may not fit within the binding site. Activity may require the simultaneous presence of multiple features. For this reason, classification methods applied to virtual screening must be able to handle nonlinear effects.

* Corresponding author fax: (902)494-1310; e-mail: weaver@chem3.chem.dal.ca.
[†] Department of Chemistry, Queen's University.
[‡] Department of Pathology, Queen's University.
[§] Dalhousie University.

10.1021/ci034143r CCC: $25.00
Published on Web 10/21/2003

Because of difficulties encountered in developing predictive RP models for classification,[5] we have implemented a genetic algorithm-based classification method that is similar in philosophy to that of RP but replaces its incremental approach to variable selection with a best-subset approach. In multivariate analyses such as QSAR, it is well-known that incremental approaches for variable selection often find suboptimal solutions when applied to large sets of variables (e.g. stepwise multiple linear regression). This has been recognized to be a limitation of RP by others; simulated annealing[16] and artificial ant colonies[17] have been used for descriptor selection. In this method, which we designate spline-fitting with a genetic algorithm (SFGA), the GFA algorithm[18] is used to select combinations of descriptor splines that partition the compounds into active and inactive groups. This differs from the typical application of the GFA algorithm in conventional QSAR,[5,19,20] in which splines and/or higher order terms are sometimes added to linear terms to account for nonlinear behavior in a series of compounds having a continuous distribution of activities. The method is implemented in Tcl scripting language for use with the Cerius2 molecular modeling package. To validate the method, we developed classification models for five data sets and compare these to the models obtained with SIMCA and RP.

METHODS

Data Sets. Data sets obtained from various sources have been used for developing classification models. The five sets are described below, ordered by increasing "HTS-likeness". An HTS-like set contains greater structural diversity than a traditional QSAR-like set and contains fewer actives than inactives. The data sets are given in the Supporting Information, both as tables of compounds and in electronic form (except for the MAO set).

(i) COX-2. A set of 467 cyclooxygenase-2 (COX-2) inhibitors has been assembled from the published work of a single research group, with in vitro activities against human recombinant enzyme expressed as IC_{50} values ranging from 1 nM to >100 μM (53 compounds have indeterminate IC_{50} values). A 314 compound subset of these inhibitors has been studied with QSAR and classification by Kauffman and Jurs.[4] They used $pIC_{50} = 6.5$ as the threshold for classifying compounds as active or inactive, guided by a histogram plot of compound counts vs pIC_{50}. Here, we employ the same threshold.

(ii) BZR. A set of 405 ligands for the benzodiazepine receptor (BZR) has been assembled mostly from the work of Haefely et al. and Cook et al. No differentiation of agonists, antagonists, and inverse agonists is made. In vitro binding affinities as measured by inhibition of [^{3}H] diazepam binding are expressed as IC_{50} values, ranging from 0.34 nM to >70 μM (65 compounds have indeterminate values). We have selected $pIC_{50} = 7.0$ as the threshold for activity by considering a histogram plot of compound counts vs pIC_{50} and the resulting balance of active and inactive compounds.

(iii) DHFR. A set of 756 inhibitors of dihydrofolate reductase (DHFR) has been assembled from the work of Queener et al. In vitro activities for *P. carinii* DHFR are reported as IC_{50} values for the inhibition of the enzymatic reduction that converts dihydrofolate to tetrahydrofolate. The IC_{50} values range from 0.034 nM to >1000 μM (83 compounds have indeterminate values). We have selected $pIC_{50} = 6.0$ as the threshold for activity.

(iv) ER. A set of 1009 estrogen receptor (ER) ligands has been assembled from multiple sources. A compilation of binding affinities for 616 nonredundant compounds has been prepared by the National Toxicology Program at the National Institute of Environmental Health Sciences. The data have been reported using the relative binding affinity (RBA) scale, which measures affinities with respect to β-estradiol. A further 393 compounds of pharmaceutical interest were culled from the chemical literature. The activity threshold RBA = 1 (after rounding to the nearest integer) was selected for designating compounds as active or inactive, a value that is useful for toxicological prioritization rather than pharmaceutical screening.

(v) MAO. A set of 1641 monoamine oxidase (MAO) inhibitors has been obtained from the RP demo distributed with Cerius2, a set that has been analyzed with other classification methods.[11,14,21] Activities are classified from 0 (inactive) to 3 (highly active). Compounds with activities of 1−3 have been grouped together into the active class. The demo file provides only precomputed descriptors and MACCS keys,[22] not the structures of the compounds.

When converting continuous activities or activity ranges into active and inactive groups, the selection of thresholds is arbitrary. The thresholds should not be viewed as parameters for optimizing models; rather the choice of thresholds should be determined by practical considerations. It may depend on the quality and quantity of leads discovered so far, with lower thresholds used at the early stages of a screening project. We did not investigate the effect of thresholds on classification models.

Computational Methodology. This work was performed mostly with the molecular modeling package Cerius2 version 4.6 (Accelrys Inc.: San Diego, CA), automated with Tcl scripts.

(i) Training and Test Set Assembly. It was the objective of this work to simulate a real virtual screening process. To this end, it was necessary to eliminate redundancy in the data sets, as virtual screening is usually applied to libraries of diverse compounds. Using 2-D (structural) fingerprints with the Tanimoto coefficient[23] (T_c) to calculate fingerprint similarity, a subset from each data set was selected using the coverage-based diversity algorithm[24,25] implemented in Cerius2. This gave sets for which all pairs of compounds have T_c values that fall below a selected threshold. For the ER and MAO sets, we used $T_c = 0.85$, a value that is often recommended for virtual screening.[26] Because of the greater similarity within the other sets, it was necessary to select higher values to retain a sufficient number of compounds. The excluded compounds were not used in any model development or evaluation.

The remaining compounds were divided between training and test sets. Approximately 40% were selected by "cherry picking" with a maximum dissimilarity algorithm[27,28] and assigned to the test set, with the remaining 60% assigned to the training set. The sets were structured this way to examine the predictive accuracy of classification methods when extrapolating outside the training set. The maximum dissimilarity algorithm (the MaxMin function in Cerius2) maximizes the minimum squared distance from each com-

Table 1. Number of Compounds, Average Tanimoto Coefficients, and Descriptors Used for Data Sets

	COX-2	BZR	DHFR	ER	MAO
before reduction	467	405	756	1009[d]	1641
T_c threshold for reduction	0.95	0.95	0.90	0.85	0.85
after coverage-based reduction	303	306	393	446	1366
train set (1/0)[a]	178 (87/91)	181 (94/87)	233 (84/149)	266 (110/156)	816 (132/684)
test set (1/0)[a]	125 (61/64)	125 (63/62)	160 (42/118)	180 (70/110)	550 (100/450)
$\langle T_c \rangle$ in train/test sets[b]	0.46/0.43	0.34/0.32	0.38/0.38	0.23/0.22	0.25/0.29
$\langle T_c \rangle$ for most similar train set-test set pair[c]	0.86	0.80	0.80	0.67	0.66
descriptors used for deriving models	35	36	33	36	38

[a] Actives are identified as 1, inactives as 0. [b] The average value of T_c between all pairs of compounds, calculated separately for the training and test sets. [c] The average value of T_c between each test set compound and the most similar training set compound. [d] There are duplicate compounds between the NTP and literature compilations.

pound to all other compounds in the selected subset, with pairwise distances determined using 1-T_c. The optimization uses a Monte Carlo procedure[28] that we have coupled to a simulated annealing protocol implemented in Tcl (up to 100 000 trial sets per pseudotemperature, which is lowered in 10% increments from 5000 to 10 K). For the ER and BZR sets, this procedure gave very different proportions of actives and inactives between the training and test sets (the more structurally diverse inactives were concentrated in the cherry-picked test sets). For those sets and all reduced training sets (see below), the MaxMin function was optimized under restraint, with a penalty applied when the ratio of actives and inactives in the selected subset differs from the ratio calculated over the complete set. The composition of sets at all stages of their preparation is summarized in Table 1.

(ii) Descriptor Generation. A number of "traditional" descriptors were calculated from the atomic composition (e.g. molecular weight) and graph-theoretic representations of molecules (e.g. χ indices,[29] E-state indices[30]). In total, 90 descriptors were calculated (the combichem defaults and E-state indices). Some descriptors were removed by examining each data set separately. The first reduction eliminated descriptors having the same value for more than 90% of compounds. The second reduction eliminated one descriptor from each pair having a pairwise correlation coefficient r satisfying $|r| > 0.95$. The number of remaining descriptors is indicated in Table 1.

(iii) SIMCA Model Development. The SIMCA method[8,9] applies principal component analysis (PCA) separately to each class of objects and uses the principal components (PCs) to define (hyper)volumes in the descriptor space. Classification of test objects is achieved by comparing the orthogonal projection distance to each class model with the PCA residuals of objects within each class. The Sybyl implementation (Tripos Inc.: St. Louis, MO) used in this work differs somewhat, as discussed in a recent evaluation of its application in drug design.[31] Descriptors were autoscaled, and no column filtering was used. The use of internal leave-one-out cross-validation in Sybyl produced models with as many components as input descriptors. Therefore, we varied the number of components systematically from 2 to 15.

(iv) RP Model Development. The RP method categorizes objects by deriving a binary decision tree in which descriptors are used to split the data set into smaller, homogeneous subsets. We have used the CART algorithm[10] implemented in Cerius2. Trees were developed using the gini splitting rule, trying 40 evenly spaced splits per descriptor, and were split until terminal nodes contained at least 2.5%, 5%, 10%, and 15% of training set compounds. The active and inactive classes were given equal weight in determining misclassification costs. Scaled pruning factors of 1−5 incremented by 0.5 were used when working with the full training sets but reduced when using subsets of the full training set (i.e. during cross-validation and training set reduction; see below). The scaled pruning factor used in Cerius2 is the complexity parameter α in ref 10 multiplied by the number of compounds in the training set. The reduction of the scaled pruning factor when using smaller sets was done such that α remains independent of training set size. Using the same scaled pruning factor instead of the same α value gives smaller trees when using subsets of the full training set. All parameters combinations were examined systematically.

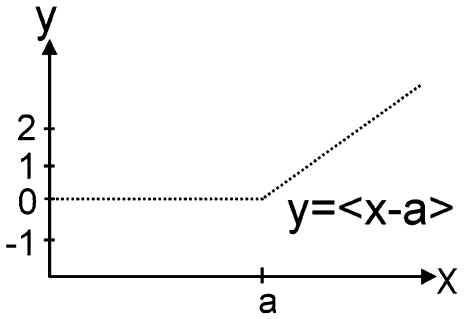

Figure 1. The truncated power spline $y = \langle x - a \rangle$.

(v) SFGA Model Development. Spline-fitting with a genetic algorithm (SFGA) uses the GFA algorithm[18] to select combinations of descriptor splines for fitting activities expressed in binary form. As for RP trees, splines partition the data set into groups having similar features and can account for nonlinear behavior. The truncated power spline $\langle x - a \rangle$ equals zero if the value of $x - a$ is negative; otherwise, it equals $x - a$ (Figure 1). The descriptor splines are selected using crossover, mutation, and knot-shift operations. The algorithm can use the same descriptor for multiple splines, enabling it to capture complex nonlinear behavior. A SFGA model has the general form

$$A_i = b_0 + \Sigma_k b_k \langle \phi_{i,k} - a_k \rangle \quad (1)$$

where A_i is the activity of compound i, $\phi_{i,k}$ is the value of descriptor k for compound i, a_k is the spline knot. The constant b_0 and coefficients b_k are determined by least-squares fitting, in which the active and inactive classes are given equal weight.[32] This is required to ensure that the models are balanced in their predictive accuracy for each category. Activities returned by a SFGA model are continu-

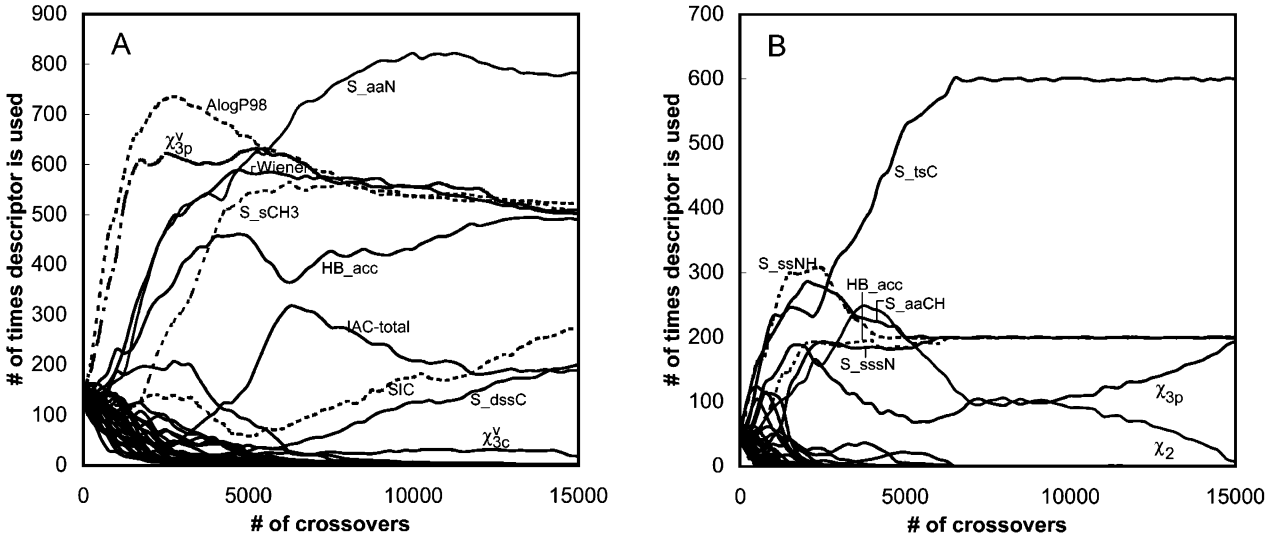

Figure 2. Frequency at which descriptors occur in the population vs the number of crossover operations performed for (A) the DHFR set and (B) the MAO set.

ous in nature. They are converted to discrete values by designating as active (1) those with predicted activities ≥ 0.5 and inactive (0) those with predicted activities <0.5. Fixed-complexity models having between 2 and 15 splines were evolved separately (we have found that the lack-of-fit fitness function in GFA is not suitable for automatically identifying the optimal number of splines). The number of crossovers, population size, mutation probability, and knot-shift probability are parameters required by the GFA algorithm. For all but the MAO set, we used 15 000 crossovers with a population of 500. For the MAO set, we used the less expensive parameters of 10 000 crossovers with a population of 200. The mutation and knot shift probabilities were set to 10% and 100%, respectively. These values were selected by systematic variation for models with eight splines generated from the DHFR and MAO sets. The combination of low mutation and high spline shift probabilities lead to populations with greater average fitness at 10 000 crossovers. This seems to achieve a reasonable balance between rapid maturation of the population (which low mutation and knot shift probabilities enhance) and maintaining sufficient diversity (which high mutation and knot shift probabilities enhance). In practice, the predictive accuracy of models does not depend strongly on the mutation and knot shift probabilities: the fittest individual has usually been found by about 5000 crossovers, and it is the only model that is retained at the end of evolution. The use of 10 000 crossovers can be seen to be more than adequate when using these parameters (Figure 2).

Run times for SFGA are roughly 30 times longer than for SIMCA and RP. For the MAO set, the development of the optimal SIMCA, RP, and SFGA models requires 0.59, 0.52, and 17.1 min on a R14K-400 MHz processor. The smaller sets require 5−8 min when using the more expensive GFA parameters. Once the model is developed, predictions are substantially faster with SFGA (0.1 s per 1000 compounds compared to 33 and 11 s for SIMCA and RP, respectively).

RESULTS

Cross-validation and the use of test sets are two approaches for assessing the ability of a model to generalize.[33] The optimal complexity of models was assessed using 10 cycles of leave-20%-out cross-validation (CV). Model complexity refers to the number of components, splits, and splines used by the SIMCA, RP, and SFGA methods, respectively. For each cycle, the training set was divided into five groups that were used in turn as prediction sets for models derived with the other four groups. The selection of descriptors was repeated for each CV model (referred to as "full" cross-validation). In addition to CV, we used the test set accuracy of models to determine the optimal complexity. We stress this amounts to nothing more than comparing a few *final* models having different feature counts, and that the test set compounds have no influence on the development of these models. The measure of predictive accuracy used in classification is the *classification rate* or percentage of compounds from each class that are correctly classified. These are denoted by C_1 for the active class (true positive rate, or *sensitivity)* and C_0 for the inactive class (true negative rate or *specificity*).

Typical plots of classification rates vs complexity are shown in Figure 3 for the RP and SFGA methods applied to the BZR set. We use two criteria to identify the optimal complexity: the number of features at which the average $\langle C_1, C_0 \rangle$ is maximized, and the number of features at which $\min\{C_1, C_0\}$ is maximized. For CV, the term "maximized" must sometimes be replaced with an arbitrary rule because the classification rate increases continuously as it approaches its high complexity value (e.g. Figure 3D). This is typical of s_{CV} vs complexity observed in conventional QSAR. Here, we use the permissive value of 1% per feature. Admittedly, Figure 3D is the most ambiguous of the (successful) CV runs; in many cases there was a clear maximum that could be used to identify the optimal number of features. In other cases, especially for the RP CV results (e.g. Figure 3B), it was not possible to identify the optimal complexity as $\langle C_1, C_0 \rangle$ and $\min\{C_1, C_0\}$ were either constant or decreased with increasing number of features.

The use of cross-validation and (large) test sets has been found to give equally reliable estimates of the true predictive accuracy of models, but the use of test sets was deemed wasteful for small data sets.[33] Consistent with those results,

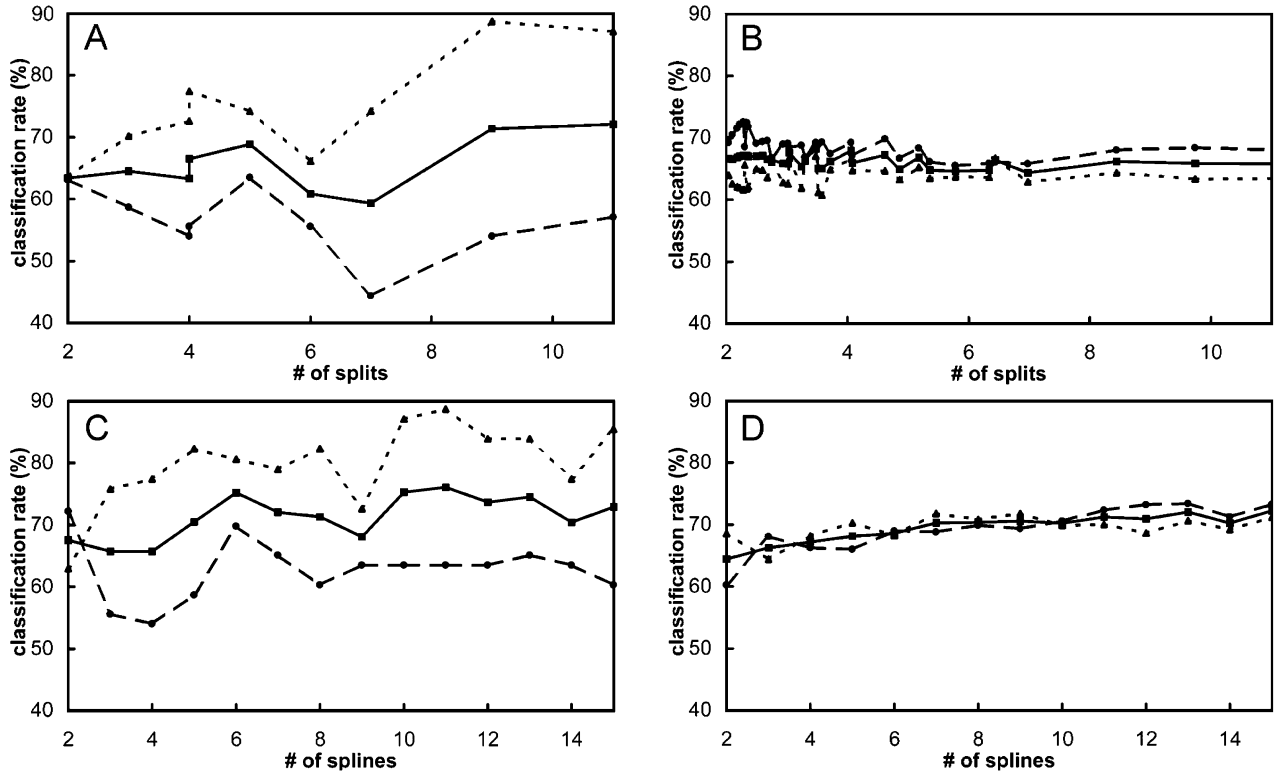

Figure 3. Classification rates of actives C_1 (dashed lines), inactives C_0 (dotted lines), and $\langle C_1, C_0 \rangle$ (solid lines) for the BZR data set. Values correspond to (A) RP and designed test set, (B) RP and cross-validation sets, (C) SFGA and designed test set, and (D) SFGA and cross-validation sets.

Table 2. Identification of Optimal Model Complexity Using Test and CV Classification Rates[a]

	SIMCA		RP		SFGA	
	test	CV	test	CV	test	CV
COX-2	**6**/6	4/6	**8**/7	-/-	**6**/6	5/5
BZR	**6**/5	7/7	**5**/5	-/-	**6**/6	8/7
DHFR	**12**[b]/12	9[a]/9	**6**/9	-/4	**6**/6	7/7
ER	5/**6**[c]	5/4	**3**/3	4/4	**3**/4	4/4
MAO	**5**/5	-/-	**4**/7	-/4	**6**/6	-/-

[a] *x*/*y*: *x* features maximize $\langle C_1, C_0 \rangle$, *y* features maximize min$\{C_1, C_0\}$; - indicates that the classification rate was flat or decreased with increasing features. The selected complexity is indicated in bold. [b] Twelve components were selected as the random test set C_0 (see below) is significantly greater for 12 than 9 components. [c] Six components were selected because the decrease in $\langle C_1, C_0 \rangle$ was small.

we have found that the optimal number of features determined using either CV or the test set is very similar (Table 2). Interestingly, the optimal number of features is generally similar among the three classification methods. Because of the large difference in computing time between test set validation and CV (50-fold in this work), the former may be preferable when analyzing large data sets with CPU-intensive methods. In this work, using the test set classification rates instead of the corresponding CV rates leads to more prominent maxima, making the identification of optimal complexity more intuitive. It has been suggested that this reflects the particular composition of the test set. While this could be true if one used a randomly assembled test set, the use of series designed to ensure that it effectively spans descriptor space minimizes this possibility. The greater "noise" in classification rates vs complexity for SFGA (Figure 3C) compared to RP (Figure 3A) or SIMCA (not shown) results from the stochastic nature of the GFA algorithm.

Classification statistics for the optimal models are summarized in Table 3; the models are given in the Supporting Information. The training set classification rates are mostly similar. For the DHFR and ER sets, the active and inactive classes are poorly fit by RP and SIMCA, respectively. SIMCA fits the MAO set less well for both activity classes. The SFGA method yields models with higher test set classification rates for all but the DHFR data set, for which SIMCA performs better (Figure 4A). For all but the BZR set, the (test set) difference between C_1 and C_0 is substantially smaller for SFGA than for SIMCA and RP. RP gives the lowest and most unbalanced test set classification rates for four data sets.

To investigate the stability of models when varying the composition of the training set, compounds were randomly assigned to training and test sets, with each set containing the same number of compounds as the corresponding designed sets. This was repeated 50 times, giving 50 pairs of random training and test sets. For each random training set, a classification model was developed using the optimal parameters.[34] Both the training and test set classification rates show the greatest variability for RP and the smallest variability for SIMCA (Table 3, Figure 4B). SFGA has the highest classification rate on the random test sets, except for the COX-2 and BZR sets for which SIMCA performs better. Trends in CV classification rates and their variability are similar (Table 3). It is important to note that the same sequence of (pseudo) random sets was used for all methods, including those used in CV. As such, classification rates are directly comparable.

Table 3. Classification Rates and Parameters for SIMCA, RP, and SFGA Models[a]

	COX-2	BZR	DHFR	ER	MAO
			SIMCA		
no. of compds	6	6	12	6	5
train	86/82	82/78	85/81	93/69	62/63
test	75/67	68/76	74/71	81/73	60/65
CV	79 ± 9/77 ± 9	73 ± 10/70 ± 12	57 ± 10/70 ± 9	83 ± 8/67 ± 7	54 ± 9/64 ± 5
random train	85 ± 3/80 ± 3	83 ± 3/80 ± 4	84 ± 4/77 ± 4	90 ± 4/69 ± 4	70 ± 5/63 ± 4
random test	75 ± 4/73 ± 5	75 ± 5/72 ± 6	62 ± 8/69 ± 6	83 ± 5/64 ± 5	58 ± 6/61 ± 4
			RP		
prune factor/min. samples	2.0/5	3.0/5	3.5/12	3.5/7	3.5/41
splits	8	5	6	3	4
train	91/81	82/81	69/85	89/75	67/75
test	79/63	64/74	57/73	76/74	63/72
CV	72 ± 12/67 ± 12	68 ± 12/65 ± 12	57 ± 12/65 ± 12	79 ± 9/72 ± 9	67 ± 12/60 ± 8
random train	88 ± 5/85 ± 5	80 ± 7/83 ± 7	78 ± 7/76 ± 5	87 ± 4/78 ± 5	76 ± 7/71 ± 5
random test	68 ± 9/67 ± 8	67 ± 9/66 ± 10	59 ± 9/66 ± 8	79 ± 9/71 ± 7	60 ± 8/67 ± 7
			SFGA		
no. of splines	6	6	6	3	6
train	83/87	81/76	80/75	87/76	71/67
test	75/72	70/81	71/66	77/80	71/70
CV	76 ± 9/72 ± 10	69 ± 11/68 ± 13	65 ± 11/64 ± 10	83 ± 8/70 ± 8	64 ± 10/68 ± 5
random train	83 ± 3/78 ± 4	80 ± 5/78 ± 6	76 ± 4/73 ± 3	87 ± 3/74 ± 5	71 ± 4/67 ± 6
random test	77 ± 6/70 ± 7	71 ± 9/69 ± 9	67 ± 7/66 ± 5	82 ± 5/72 ± 6	64 ± 6/66 ± 7

[a] Classification rates for actives (sensitivity) and inactives (specificity) are indicated as C_1/C_0 for the designed training and test sets and as $C_1 \pm \sigma_1/C_0 \pm \sigma_0$ for CV and random sets.

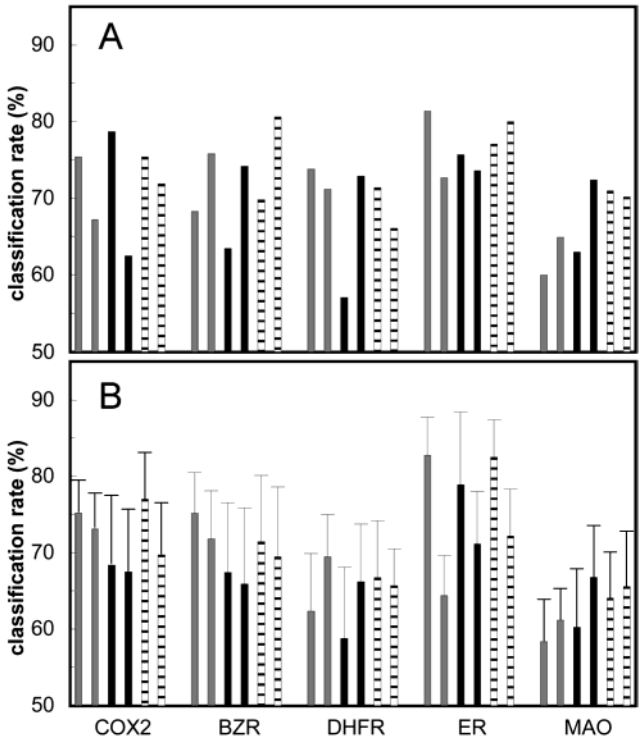

Figure 4. Classification rates for SIMCA (gray), RP (black), and SFGA (hashed) applied to five data sets. Each pair of matching bars corresponds to C_1 (left) and C_0 (right) for (A) the designed test set and (B) the average over 50 random test sets. For (B), error bars correspond to the standard deviation.

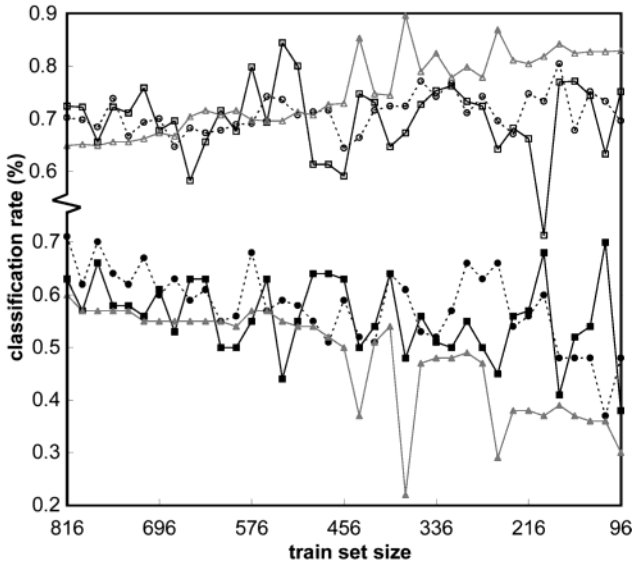

Figure 5. Effect on MAO test set classification rates C_1 (filled markers, bottom series) and C_0 (open markers, top series) from reducing the size of the training set for SIMCA (gray), RP (black), and SFGA (dotted). The top and bottom series have been separated for clarity.

Another approach for introducing training set variation consists of gradually eliminating compounds from the designed training set and assessing the effect on the designed test set (which does not vary). A series of nested subsets of compounds was assembled from the full training set using the maximum dissimilarity algorithm, as described in the methods section for test set design. The compounds to be retained were selected from the previous set, as opposed to selecting the compounds to be discarded; this serves the purpose of maintaining as much as possible an effective coverage of descriptor space while the number of compounds is reduced. For the MAO set, the full training set was decremented by 20 compounds until 96 remained. For the other sets, the full training set was decremented by five compounds (10 for ER), until about 50 remained. The optimal parameters were used for generating models. Figure 5 shows the results obtained for the MAO set. To quantify trends in the variation of classification rates with training set size, linear regression was used to fit a line to each series in Figure 5 and the corresponding series for the other data sets. The slopes (expressed as % decrease in $C_{1/0}$ per 100

Table 4. Effect of Training Set Reduction on Test Set Classification Rates[a]

	SIMCA		RP		SFGA	
	m (1/0)	*s* (1/0)	*m* (1/0)	*s* (1/0)	*m* (1/0)	*s* (1/0)
COX-2	−0.8/0.4	2.4/3.1	6.8/−4.1	4.2/4.9	6.6/4.9	3.9/3.5
BZR	−6.9/11.7	3.7/3.6	−3.2/6.0	9.0/5.1	8.0/12.1	5.2/3.6
DHFR[b]	6.7/−1.1	4.3/4.3	−5.6/5.2	10.2/5.7	2.0/2.6	5.6/4.2
ER	10.0/−3.5	3.9/3.1	−3.1/4.0	4.7/4.4	−0.2/0.9	5.8/4.1
MAO	3.7/−3.0	5.8/3.3	1.1/0.1	7.2/7.2	2.1/−0.7	5.7/3.3

[a] The slope *m* from linear regression is expressed as the percent decrease in $C_{1/0}$ per 100 compounds removed from the training set; negative values indicate an increase in $C_{1/0}$; *s* is the standard error of regression; *x*/*y* correspond to the active/inactive classes. [b] Sets with 58 and 53 compounds were excluded, as SIMCA and RP had extremely low values of C_1.

compounds removed from the training set) and standard error of regression about the fit lines are given in Table 4.

By comparing the standard errors of regression, it emerges that SIMCA classification rates are the least variable upon training set reduction, and RP classification rates are the most variable by a wide margin. The slopes indicate that SIMCA suffers the smallest decrease in classification rates. One might also conclude that the RP classification rates decrease less than those for SFGA. The increases seen for the COX-2 inactives, and especially the BZR and DHFR actives, are coupled to large standard errors; those series are not well-described by linear regression. Indeed, if only the largest 11 of 27 sets are considered for BZR, C_1 decreases by 58% per 100 compounds removed for RP. The RP slopes with lower standard errors are similar to those for SFGA. On average, both methods appear to be similarly affected by training set reduction.

In addition to examining the effect of training set variation on classification rates, it is instructive to compare the consistency of model features (i.e. the descriptors that are selected). Because SIMCA models are defined using latent structures, the top *x* descriptors ranked by "discriminating power" were selected for comparisons with RP and SFGA, where *x* corresponds to the number of components included in the model. For the 50 models corresponding to the 50 randomly assembled training sets, the number of occurrences of each descriptor was determined and divided by the total number of features present in the 50 models (e.g. 50 × 6 for six feature models). The consistency of models must be considered in light of the correlations among the descriptors, as it is reasonable to expect that highly correlated descriptors will be used interchangeably. A qualitative picture of model consistency can be obtained by applying principal component analysis[7] to the autoscaled descriptor matrix and representing the frequency of occurrence of descriptors in the loading space corresponding to the first 2 or 3 principal components (Figure 6A). Proximity of descriptors within this space indicates that they represent similar molecular properties. A classification method that produces consistent models will give a limited number of descriptors or descriptor clusters that account for most of the features from the 50 models. In Figure 6A, this corresponds to large circles or clusters of intermediate circles. A method which produces less consistent models will yield smaller circles distributed throughout the loading space. It can be seen that for the MAO random sets, SIMCA produces the most consistent models followed by SFGA and RP.

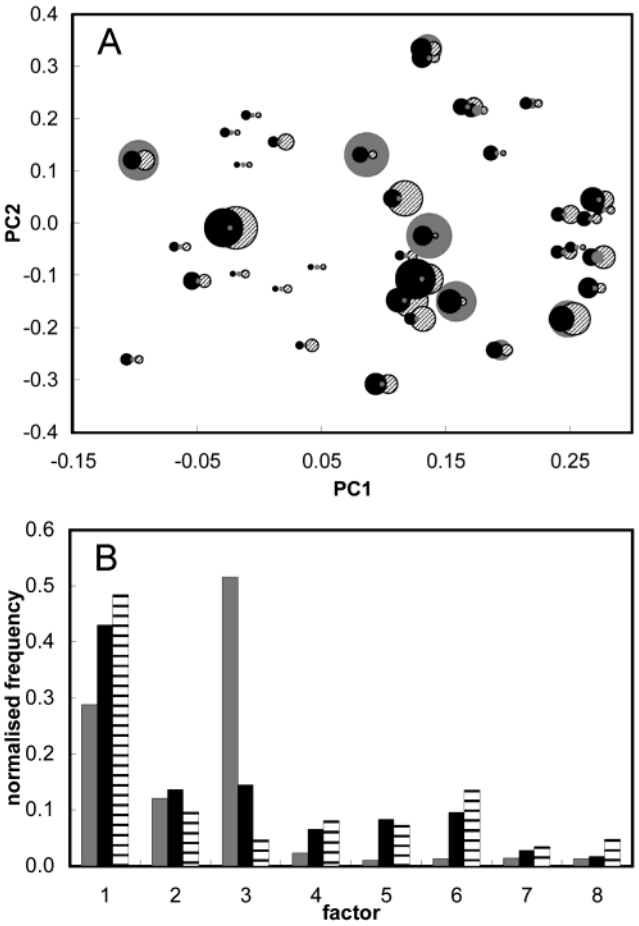

Figure 6. Variability of descriptors in 50 MAO random training set models from SIMCA (gray), RP (black), and SFGA (hashed). (A) The 38 descriptors used for developing models represented in the loading space of the first two principal components. The area of circles is proportional to the frequency at which the descriptor occurs among the 50 models. The circles for RP and SFGA have been shifted by −0.05 and +0.05 for clarity. (B) Descriptor frequencies converted to eight varimax-rotated factors. See text for details.

The consistency of the models was quantified using the following procedure. Factor analysis[7] was used to obtain eight varimax-rotated factors from the autoscaled descriptor matrix; eight factors account for roughly 80% of its variance. Varimax-rotated factors differ from PCA factors in that the rotated factors have high correlations with one smaller set of descriptors and little or no correlation with another set of descriptors, making the factors more interpretable and distinct from each other. For each factor, the fraction of a descriptor's variance that it explains was multiplied by the descriptor's observed frequency; the products between a given factor and all descriptors were summed and divided by the total for all eight factors. The final result can be viewed as observed descriptor frequencies expressed compactly in terms of eight independent factors. A histogram of factor frequencies leads to similar conclusions for the MAO set (Figure 6B). A set of maximally inconsistent models would have equal frequencies for all eight factors, while a set of consistent models would have higher frequencies for certain factors. Figure 6B can therefore be summarized by calculating the χ^2 goodness-of-fit statistic for SIMCA, RP, and SFGA, using the null hypothesis of equal factor frequencies. A set of consistent models will have a high value of χ^2. The same procedure

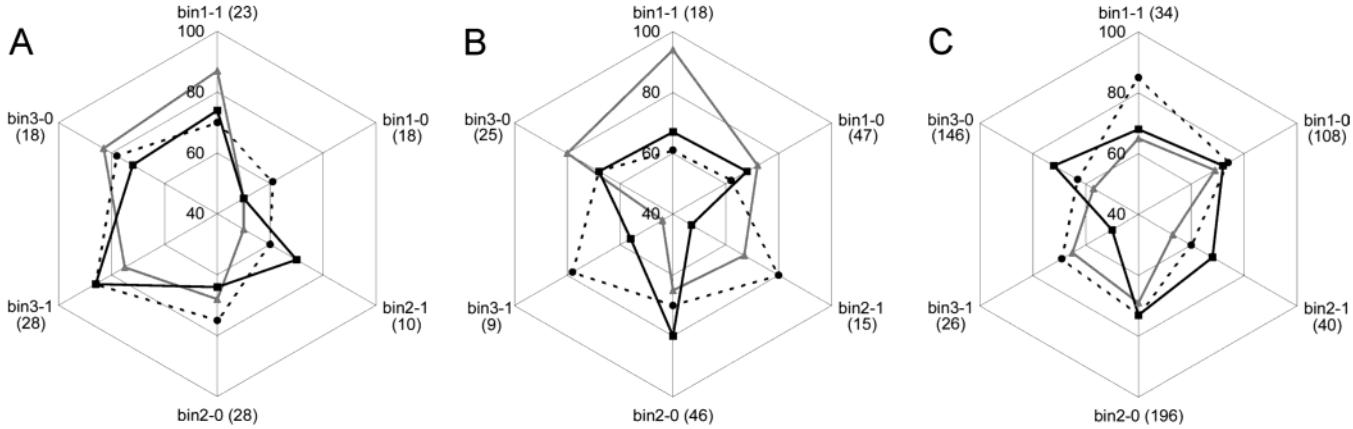

Figure 7. Graphical representation of test set classification rates C_1 and C_0 from SIMCA (gray), RP (black), and SFGA (dashed) in each of the three distance bins that describe the extent of extrapolation from the training set. The axis label "bin1-0" correspond to C_0 for distance bin 1; similar definitions apply to the other five axes. Numbers in parentheses indicate the number of compounds of a given activity class in each bin. (A) COX-2 (0.09, 0.15), (B) DHFR(0.18, 0.24), and (C) MAO (0.27, 0.38); the thresholds used for defining bins are indicated.

Table 5. χ^2 Representation of Descriptor Consistency among Models from Random and Reduced Training Sets

	random training sets χ^2			reduced training sets χ^2		
	SIMCA	RP	SFGA	SIMCA	RP	SFGA
COX-2	0.66	0.23	0.42	0.70	0.32	0.72
BZR	0.99	0.50	0.76	2.39	0.52	0.69
DHFR	0.87	0.78	0.91	1.21	0.73	1.84
ER	0.67	0.44	0.92	0.73	0.62	0.87
MAO	1.92	0.97	1.24	1.81	1.44	1.88

was repeated using models obtained from training set reduction, with results given in Table 5. It should be noted that this procedure would not be appropriate if the individual models actually contained eight completely orthogonal features; a set of identical models would then have a low value of χ^2. The use of eight factors for representing the descriptors is a reasonable compromise between representing the descriptor pool more accurately and maintaining sufficient "occurrences" for each factor to allow statistics such as χ^2 to be applicable.

For the random training sets, it emerges that SIMCA produces the most consistent models for 3 sets, while SFGA produces the most consistent models for the other 2 sets. For training set reduction, SFGA produces the most consistent models for all but the BZR set. There is a higher degree of correlation among the reduced training sets than the random sets, and the difference between χ^2 values for SIMCA and SFGA tend to be greater when SIMCA produces the larger value. Thus, it is clear than SIMCA is generally more consistent than SFGA when training set variation is extensive. In all cases, RP produces the least consistent models.

The predictive accuracy of the classification methods when extrapolating from the training set was examined. The values of $1\text{-}T_c$ were calculated between each test compound and all training compounds (for the designed sets), and the smallest value was used to represent the compound's distance from the training set. The distances were binned into three intervals, and test set classification rates were calculated for each (Figure 7). In this representation, each pair of axes represents C_1 and C_0 in one distance bin; a method which gives balanced values of C_1 and C_0 that are similar across all distance bins will give a hexagonal shape; larger "hexagons" indicate greater predictive accuracy. In general, SFGA gives more balances values of C_1 and C_0, even as distance to the training set increases.

Table 6. Classification Rates Obtained from a Consensus of SIMCA, RP, and SFGA Models

	consensus C_1/C_0	% of compds with prediction (1/0)
COX-2	90/71	69/77
BZR	76/92	65/63
DHFR	89/87	45/47
ER	93/84	64/74
MAO	72/79	76/72

While it is important to assess differences between the classification methods, it should be noted that they are most powerful when applied in concert. Substantially improved classification rates can be obtained when using a unanimous agreement consensus scheme, in which all three methods must agree on the activity of a compound for a prediction to be registered (Table 6). Consequently, predictions are not available when the methods do not agree on a predicted activity. This should not be a severe limitation. In situations for which the number of compounds to be synthesized and tested must be minimal, then only the consensus-predicted actives could be screened. On the other hand, if the risk of missing actives must be minimized, then only those compounds that are consensus-predicted inactive could be excluded. A consensus scheme using a majority agreement (i.e. 2 actives, 1 inactive = active) would give predictions for all compounds. Unfortunately, this gives classification rates intermediate to those obtained from the most and least predictive models. It does, however, give classification rates slightly greater than the average among the three methods.

DISCUSSION

In this paper, we have compared soft independent modeling by class analogy (SIMCA), recursive partitioning (RP), and spline-fitting with a genetic algorithm (SFGA) in their classification performance for five data sets. The use of latent variables for defining SIMCA classification models results in classification rates and model features that are insensitive to variations in the training set. The addition or elimination of compounds may significantly alter the profile of values for some descriptors, but the underlying latent variables are

less affected because of their relationship to multiple descriptors. Some limitations of SIMCA can be deduced from the present work. First, SIMCA cannot handle distinct clusters of active compounds embedded within a larger space of inactive compounds. This may be partly responsible for the lower classification rates observed for the MAO set, as the diverse group of active compounds may elicit their biological response through multiple binding modes. A limitation in the Sybyl implementation of SIMCA is the inability to fix separately the number of components used for each category. This can be observed in the variation of the classification rates C_1 (for actives, sensitivity) and C_0 (for inactives, specificity) with the number of components for the ER and DHFR sets (data not shown). For ER, C_1 is maximized at three components, although C_0 requires six components. A similar situation occurs for DHFR, where C_1 and C_0 are maximized at 7 and 12 components, respectively. It should be possible to vary the components separately in Sybyl by implementing a "custom" SIMCA using SPL and the factor analysis routine. Alternatively a model could be derived for the active class alone, and compounds falling outside the active model are assigned to the inactive class.[9]

RP produces models that are more interpretable than SIMCA models, because of their intuitive treelike structure using the original descriptors. It can be applied to extremely large data sets by virtue of its speed.[35] The short run times of RP derive from its use of an incremental approach for choosing descriptor splits. However, it is responsible for RP's sensitivity to the composition of the data set. Once a split has been made, it cannot be changed later during tree growth even if another descriptor would achieve a higher classification rate. This limitation is less acute for simple models, as shown by the performance of RP on the ER data set. Attempts to improve on the incremental approach invariably cause increases in the time required for deriving models (e.g. optimizing descriptor combinations by simulated annealing takes 30 times longer than traditional RP[16]).

SFGA represents our attempts to retain RP's partition-based approach to classification while replacing its incremental descriptor selection with a best-subset selection. In general, the predictive accuracy of SFGA models developed in this work exceeded those of SIMCA and RP models. Using a genetic algorithm yields models with greater consistency of features and stability of classification statistics than the corresponding RP models when the composition of the training set is varied. The SFGA classification rates C_1 and C_0 are also more balanced than those from SIMCA and RP. The run times for SFGA are substantially longer. It is possible to decrease it by using only 5000 crossover operations, as suggested by Figure 2. Despite longer run times, a complete set of models can be developed in 2−3 h on a typical workstation. With the model at hand, SFGA is substantially faster at making predictions. Nevertheless, the short times required for developing SIMCA and RP models may warrant their use over more time-consuming methods when developing preliminary models. The SFGA classification rates are only moderately higher in most cases.

A comparison of classification methods is inherently dependent on the descriptors employed. Pearlman and Smith have discussed the limitation of traditional molecular descriptors applied to the active vs inactive problem.[36] To some extent, the present study validates their point of view. This is most evident in the moderate increase in *training* set classification rates as the number of model features increases. For the MAO set, the difference between the highest and lowest values of $\langle C_1, C_0 \rangle$ amounts to 8%, 11%, and 7% for SIMCA, RP, and SFGA, despite large increases in complexity by 12−18 features. Despite the limitations of traditional 1-D and 2-D descriptors, they remain useful for developing classification models and confer the important advantage of requiring no structural optimization of molecules. High-dimensional descriptors such as molecular fingerprints (excluding low-complexity variants such as the 166 MACCS keys[22] and minifingerprints[37]) that are often used in clustering are clearly not applicable for developing classification models. The BCUT metrics[36] and the interclass distance parameter[38] represent recent efforts for developing descriptors more suitable for classification. However, the use of BCUT metrics with the binary QSAR method[13,14] produced models of predictive accuracy similar to those obtained with traditional descriptors.[39]

CONCLUSIONS

Because of the shift toward combinatorial chemistry and high-throughput screening in drug discovery, the use of classification methods is likely to continue increasing. The present work further establishes their usefulness in screening, for which efficiency can be increased by directing physical resources to those compounds predicted to be active. We have described spline fitting with a genetic algorithm (SFGA), a method that uses descriptor splines to partition compounds into active and inactive classes. SFGA was compared to soft independent modeling by class analogy (SIMCA) and recursive partitioning (RP), two well-established classification methods, by using five data sets designed to maximize their diversity. SFGA produced the most predictive models for four of five designed test sets. Similar results were obtained when using sets assembled by random selection. The stability of SFGA classification statistics upon training set variation was found to be intermediate to those of RP and SIMCA, with SIMCA giving the most stable model statistics. A similar trend was observed with respect to the features used in the models. Despite the inevitable rankings that such comparisons produce, we showed that a consensus approach involving all three classification methods outperforms the best single method in all cases.

ACKNOWLEDGMENT

D.F.W. acknowledges support from the Natural Sciences and Engineering Research Council (NSERC), the Canadian Institutes of Health Research (CIHR), and the Canada Research Chairs Program. J.J.S. acknowledges support from a CIHR doctoral research award.

Supporting Information Available: Data sets in tabular and electronic form (MDL SD format), including classification activities, set membership and literature references; classification models, and the Tcl script for implementing SFGA in Cerius2. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Walters, W. P.; Stahl, M. T.; Murcko, M. A. Virtual Screening − an Overview. *Drug Discov. Today* **1998**, *3*, 160−178.

(2) Dean, P. M. *Molecular Similarity in Drug Design*; Blackie Academic: London, 1995.
(3) Martin, Y. C.; Brown, R. D.; Bures, M. G. Quantifying Diversity. *Combinatorial Chemistry and Molecular Diversity in Drug Discovery*; Wiley: New York, 1998; pp 369−385.
(4) Kauffman, G. W.; Jurs, P. C. QSAR and K-Nearest Neighbor Classification Analysis of Selective Cyclooxygenase-2 Inhibitors Using Topologically-Based Numerical Descriptors. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1553−1560.
(5) Sutherland, J. J.; Weaver, D. F. Development of Quantitative Structure−Activity Relationships and Classification Models for Anticonvulsant Activity of Hydantoin Analogues. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1028−1036.
(6) Bajorath, J. Selected Concepts and Investigations in Compound Classification, Molecular Descriptor Analysis, and Virtual Screening. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 233−245.
(7) Everitt, B. S.; Dunn, G. *Applied Multivariate Analysis*; Oxford University Press: New York, 1992.
(8) Wold, S. Pattern-Recognition by Means of Disjoint Principal Components Models. *Pattern Recognition* **1976**, *8*, 127−139.
(9) Dunn III, W. J.; Wold, S. Simca Pattern Recognition and Classification. *Chemometric Methods in Molecular Design*; VCH: New York, 1995; pp 179−193.
(10) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees*; Wadsworth: Belmont, CA, 1984.
(11) Hawkins, D. M.; Young, S. S.; Rusinko, A. Analysis of a Large Structure−Activity Data Set Using Recursive Partitioning. *Quant. Struct.-Act. Relat.* **1997**, *16*, 296−302.
(12) Godden, J. W.; Furr, J. R.; Bajorath, J. Recursive Median Partitioning for Virtual Screening of Large Databases. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 182−188.
(13) Gao, H.; Williams, C.; Labute, P.; Bajorath, J. Binary Quantitative Structure−Activity Relationship (QSAR) Analysis of Estrogen Receptor Ligands. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 164−168.
(14) Gao, H.; Lajiness, M. S.; Van Drie, J. Enhancement of Binary QSAR Analysis by a GA-Based Variable Selection Method. *J. Mol. Graph.* **2002**, *20*, 259−268.
(15) Dixon, S. L.; Villar, H. O. Investigation of Classification Methods for the Prediction of Activity in Diverse Chemical Libraries. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 533−545.
(16) Blower, P.; Fligner, M.; Verducci, J.; Bjoraker, J. On Combining Recursive Partitioning and Simulated Annealing to Detect Groups of Biologically Active Compounds. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 393−404.
(17) Izrailev, S.; Agrafiotis, D. A Novel Method for Building Regression Tree Models for QSAR Based on Artificial Ant Colony Systems. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 176−180.
(18) Rogers, D.; Hopfinger, A. J. Application of Genetic Function Approximation to Quantitative Structure−Activity-Relationships and Quantitative Structure−Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854−866.
(19) Shi, L. M.; Fan, Y.; Myers, T. G.; O'Connor, P. M.; Paull, K. D. et al. Mining the NCI Anticancer Drug Discovery Databases: Genetic Function Approximation for the QSAR Study of Anticancer Ellipticine Analogues. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 189−199.
(20) Fan, Y.; Shi, L. M.; Kohn, K. W.; Pommier, Y.; Weinstein, J. N. Quantitative Structure-Antitumor Activity Relationships of Camptothecin Analogues: Cluster Analysis and Genetic Algorithm-Based Studies. *J. Med. Chem.* **2001**, *44*, 3254−3263.
(21) Brown, R. D.; Martin, Y. C. Use of Structure Activity Data to Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.
(22) MDL Information Systems Inc.: San Leandro, CA.
(23) Willett, P.; Winterman, V. A Comparison of Some Measures for the Determination of Intermolecular Structural Similarity Measures of Intermolecular Structural Similarity. *Quant. Struct.-Act. Relat.* **1986**, *5*, 18−25.
(24) Clark, R. D. Optisim: An Extended Dissimilarity Selection Method for Finding Diverse Representative Subsets. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1181−1188.
(25) Reynolds, C. H.; Druker, R.; Pfahler, L. B. Lead Discovery Using Stochastic Cluster Analysis (SCA): A New Method for Clustering Structurally Similar Compounds. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 305−312.
(26) Matter, H. Selecting Optimally Diverse Compounds from Structure Databases: A Validation Study of Two-Dimensional and Three-Dimensional Molecular Descriptors. *J. Med. Chem.* **1997**, *40*, 1219−1229.
(27) Lajiness, M.; Johnson, M. A.; Maggiora, G. M. Implementing Drug Screening Programs Using Molecular Similarity Methods. *QSAR: Quantitative Structure−Activity Relationships in Drug Design*; Alan R. Liss Inc.: New York, 1989; pp 173−176.
(28) Hassan, M.; Bielawski, J. P.; Hempel, J. C.; Waldman, M. Optimization and Visualization of Molecular Diversity of Combinatorial Libraries. *Mol. Divers.* **1996**, *2*, 64−74.
(29) Hall, L. H.; Kier, L. B. The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure−Property Modeling. *Reviews in Computational Chemistry*; 1991; Vol. 2, pp 367−422.
(30) Hall, L. H.; Kier, L. B. Electrotopological State Indexes for Atom Types − a Novel Combination of Electronic, Topological, and Valence State Information. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1039−1045.
(31) Hunt, P. A. QSAR Using 2D Descriptors and Tripos' SIMCA. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 453−467.
(32) Without having access to the GFA source code, the only way to give both classes equal weight is to introduce duplicate entries for the minority class in the study table. If the ratio of inactives to actives is *x.y*, all actives are duplicated (*x-1*) times, and a further *0.y* × (no. of actives) active compounds are randomly selected and duplicated. This has the unfortunate consequence of making the calculations more time-consuming.
(33) Hawkins, D. M.; Basak, S. C.; Mills, D. Assessing Model Fit by Cross-Validation. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 579−586.
(34) For RP applied to the BZR, DHFR, and ER sets, the same tree was obtained for multiple values of the scaled pruning factor. The median of those scaled factors was used for generating RP models from the random and reduced training sets.
(35) Rusinko, A.; Farmen, M. W.; Lambert, C. G.; Brown, P. L.; Young, S. S. Analysis of a Large Structure/Biological Activity Data Set Using Recursive Partitioning. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1017−1026.
(36) Pearlman, R. S.; Smith, K. M. Metric Validation and the Receptor-Relevant Subspace Concept. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 28−35.
(37) Xue, L.; Godden, J. W.; Bajorath, J. Evaluation of Descriptors and Mini-Fingerprints for the Identification of Molecules with Similar Activity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1227−1234.
(38) Claycamp, H. G.; Sussman, N. B. A Simple Inter-Class Distance Parameter for Predictive SAR/QSAR Models. *Quant. Struct.-Act. Relat.* **1999**, *18*, 11−15.
(39) Gao, H. Application of BCUT Metrics and Genetic Algorithm in Binary QSAR Analysis. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 402−407.

CI034143R