A Chemical Class-Based Approach to Predictive Model Generation

David W. Miller*

Sage Informatics LLC, 825 Calle Mejia 1103, Santa Fe, New Mexico 87501

Received September 17, 2002

We make a quantitative comparison of two distinct approaches to predictive model generation in the context of diverse screening data. In the default approach, a single recursive partitioning model is constructed using all of the training data at one time. In the "class-based" approach, the same data are first partitioned into homogeneous, scaffold-based classes, and models are constructed within each class independently. Both approaches are tested on the identical set of hold-out data, using a formal protocol that includes consensus scoring to handle the multiple class-based models. The entire process is performed using three different descriptor sets and is repeated using five separate random trials, such that the trial-averaged prediction rates for the two approaches can be quantitatively compared. We find that although the predictive performances of the class-based and default approaches are similar, the former has at least two distinct advantages. The first is greater interpretability, in that chemists can more easily extract useful structure−activity information from the models. The second is greater reliability, allowing models to be applied with increased confidence to unseen data in virtual-screening applications.

INTRODUCTION

In recent years, chemical lead discovery and optimization efforts have seen a significant increase in the use of predictive learning methods borrowed from the engineering and computational sciences. Specific approaches vary widely, though the most common are either based on or incorporate major elements from linear regression equations,[1,2] neural networks,[3−5] expert systems,[6−8] tree-based approaches such as recursive partitioning,[9−11] and kernel (nearest-neighbor) methods.[11−13] The particular application areas of these methods are equally varied and include biological target inhibition,[3,4,9] cell membrane permeability,[14−16] toxicity,[6−8,17] and similarity to known drugs,[18,19] among others. Although far too numerous to cover in this brief citation list, these various computational methods and applications have been the subject of a number of recent literature reviews.[5,20−23]

Given the increasing use of high-throughput screening (HTS) approaches in drug discovery, a growing issue of concern regarding predictive methods in chemistry is the ability to construct models using diverse data. In typical HTS settings, the degree of structural homogeneity present in a set of screening data may be quite low: that is, the molecules contained in such a set might collectively represent a large number of distinct chemical families, each with a relatively unique structural origin. In such a setting, the underlying biological mechanisms that the data represent are likely to be sufficiently diverse that a single model—particularly one that is mathematically simple—cannot completely uncover the corresponding structure−activity relationships (SAR). Thus many common modeling approaches, such as multiple linear regression, the popular CoMFA technique,[24] and even many pharmacophore-elucidation methods, prove inadequate in situations where the underlying data are structurally diverse.

As a result of this limitation, there has been an increased focus on more powerful modeling techniques, particularly those having the flexibility to partition data into distinct groups and model each of them independently. Many traditional learning methods take this approach. For example, in the K nearest neighbor (KNN) and recursive partitioning (RP) methods, activity predictions are performed using localized models derived from only a portion of the training data, the first via neighborhoods centered around individual molecules and the latter via a disjoint partitioning of the entire variable (or "descriptor") space. Given an appropriate set of descriptors capable of distinguishing chemical structures, both of these methods can achieve reasonable predictive performance even in an HTS setting.[9−13,25,26]

In addition to these standard approaches, an increasing number of novel methods are being introduced that address the problem of diverse chemical data more directly, usually through an integrated organizational step. For example, the expert system MCASE[8,27] constructs localized QSAR models within groups of molecules, where the groups are defined using a list of activity-correlated fragments—called "biophores" and "biophobes"—identified dynamically from the initial data. The LeadScope program[28] takes a somewhat different approach, organizing data as a separate and independently valuable process. Here a large, preconfigured catalog of substructural elements partitions molecules into well-known chemical families, and activity correlations are identified within each family as a subsequent procedure. As yet another approach, Blower et al.[29] modify the standard recursive partitioning algorithm such that the grouping of molecules is based on combinations of, rather than single, fragments. This modification, based on a simulated-annealing procedure, is made with the express goal of increasing the structural homogeneity of the localized data subsets (here, RP nodes in the tree) from which activity patterns are subsequently extracted. What these and many other recently described learning methods have in common is the integra-

* Corresponding author phone: (505)989-7650; e-mail: dmiller@sageinformatics.com.

10.1021/ci025606g CCC: $25.00 © 2003 American Chemical Society
Published on Web 01/24/2003

tion of an explicit molecular organization process into an overall method designed for activity prediction. The increased flexibility imparted by such an organizational step allows these methods to be effectively applied to the type of inhomogeneous data common in HTS settings.

In light of the growing number of methods incorporating both organizational and predictive steps, it is important that there be a clear understanding of exactly how these two fundamental components are related to one another, and how they each contribute to overall model performance. To this end, we have devised a protocol in which two distinct approaches to predictive model generation—one in which the organizational step is lacking altogether, and another in which both steps are present but completely decoupled—are quantitatively compared. In the first ("default") approach, a single model is constructed from the complete set of data, using a standard QSAR method sufficiently complex to yield good predictive performance. In the second ("class-based") approach, the identical data are initially subjected to an independent clustering step that divides that data into groups, or "classes," on the basis of similarity only, with no regard for molecule activity; then as a subsequent step, a predictive model is constructed independently within each class, using the identical QSAR method and descriptors as in the default case. Both of these approaches are intended to represent general, broadly defined strategies, and as such, are not designed as specific new learning algorithms per se. Instead, our aim is to set forth a well-defined protocol, including cross-validation, complexity control, hold-out testing, and consensus scoring of multiple models, so that a direct and detailed comparison can be made between these two closely related procedures.

In making this comparison, we expect to derive at least two important benefits. First, the comparison will help to delineate the effect of molecular organization procedures on not only predictive performance but also other more qualitative aspects of model development. An important example of the latter is model interpretability, broadly defined as the ease with which useful structure−activity information can be inferred from the models. This is a particularly relevant issue in chemistry, where most QSAR efforts are intended not simply to perform "black box" prediction but rather to impact directly the decision making of chemists, who often use the models to guide subsequent steps in a molecular design. Second, should the class-based strategy compare favorably with the default in terms of prediction, it would suggest the possibility that many standard QSAR methods might be greatly enhanced simply through the incorporation of an initial class-generation process. Since the procedure used for this initial organization step can be selected from a wide variety of available methods, many designed specifically to facilitate data analysis by chemists, resulting models will likely have a high degree of intrinsic interpretability, and therefore an increased utility, to end users, without the need to devise a complex new learning algorithm.

In the sections that follow, we present a comparison of the default and class-based approaches using a diverse set of molecules screened against HIV. The algorithms used for model construction are described in the Methodology section and include a class-generation procedure based on the organization of molecules by common scaffold; a learning method derived from standard recursive partitioning; and a broad set of descriptor types based on structural fragments, pharmacophores, and bulk molecular properties. Each of these has been chosen to be relatively general, such that results should pertain broadly to a default or class-based strategy rather than to any particular protocol or algorithm. Also outlined in the Methodology section are protocols used for the comparison, including a training procedure based on cross-validation and complexity control to prevent overfitting; a testing procedure based on a fixed set of hold-out data to ensure unbiased results; and a performance metric, averaged over five random trials, measuring the fraction of hold-out molecules correctly predicted as being either active or inactive. In the Results section, the predictive performances of the class-based and default approaches are quantitatively compared, while the Discussion section provides further comparisons and general conclusions, with particular focus on the effect of each approach on model interpretability and virtual-screening performance.

It should be noted that all tasks involving the preparation and visualization of chemical data, the generation of structural classes and molecular descriptors, and the construction and virtual screening of QSAR models were performed using an in-house chemoinformatics software package called ChemTK.[30]

METHODOLOGY

Data Sets. Chemical and biological screening data were obtained from the National Cancer Institute (NCI).[31] The initial data set contained 249 081 molecules, each having two-dimensional (2D) coordinates and associated screening results for cancer (results as of August 1999) and/or AIDS (results as of October 1999). From this initial collection we extracted the AIDS-related subset, yielding a reduced data set of size 42 689. The biological values for these molecules are derived from a cell-based assay measuring protection from HIV-1 infection and are categorized as confirmed active (CA), confirmed moderately active (CM), or confirmed inactive (CI). Following removal of 139 molecules having nonstandard valences (e.g., pentavalent carbon atoms) and other errors, the final number of molecules in each category were 423, 1078, and 41 049, respectively. The CA- and CM-class molecules were then combined into one "active" class, yielding a final data set of size 42 550, of which 1501 were designated active and 41 049 inactive.

Model Descriptors. Because the focus of this study is to compare two distinct approaches to model generation and not to evaluate descriptors per se, we have chosen to implement three broad classes of descriptors and to perform model development using each of them separately. While our implementations are relatively simple and do not cover the breadth of descriptor types available in the literature (see recent reviews[23,32]), they are intended to be sufficiently diverse to avoid excessive bias and to ensure that our conclusions regarding the class-based approach are general ones not confined to a particular methodology or experimental design. The three descriptor classes, referred to as unbranched-fragment (**UF**), pharmacophore (**PH**), and fixed-property (**5P**) descriptors, are described below. A fourth type, based on ring systems (**RS**), was used only for the default approach and will be discussed in a subsequent section.

The first type of descriptor (UF) is based on unbranched molecular fragments ("linear paths") of between two and four

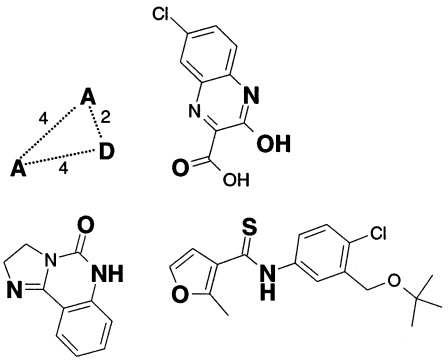

Figure 1. Three NCI molecules containing the identical graph-based pharmacophore shown at the upper left. This pharmacophore represents one hydrogen-bond donor (D) and two hydrogen-bond acceptors (A), with interfeature distances of 2 bonds (A−D), 4 bonds (A−D), and 4 bonds (A−A). Bold atoms indicate the location of the pharmacophore within each molecule.

atoms. Descriptors of this type are generated using a dynamic search procedure: for each molecule in the initial data set, all such fixed-length paths through the molecule are identified on the fly, and each is stored in a final unique list of descriptors representing the entire input set. In the end, each molecule is assigned a fixed-length binary vector in which 0/1 values indicate the absence/presence within the molecule of each fragment from the final unique list. Matching rules for atoms are based on atom symbol and aromaticity and for bonds are based on bond type only (single, double, triple, or aromatic).

The second descriptor type (PH) is based on topological, or "graph-based," pharmacophores. Each such pharmacophore consists of two parts. The first is a list of between two and three pharmacophore features, the four examples of which are hydrogen-bond donors and acceptors (where definitions are based on Greene et al.[33]) and formal positive and negative charges. The second is a list of interfeature distances. In this graph-based implementation, each distance is measured using bond connectivity rather than a typical three-dimensional distance; thus two features separated by three bonds would have an interfeature distance of 3, regardless of the relative positions of the two points in Cartesian space. Figure 1 provides a graphical illustration of this descriptor type and shows how molecules from distinct chemical classes can share an identical pharmacophore representation despite their topological dissimilarity.

As in the previous case of unbranched fragments, PH descriptors are generated dynamically to produce a final unique list of pharmacophores representing the entire initial data set. To restrict the search space, a size constraint is placed on each pharmacophore such that the number of features must be between 2 and 3, and each interfeature distance must be between 1 and 4. In cases where more than one path, and hence distance, exists between a single pair of features, all possible paths are used to generate distinct pharmacophore representations. As before, the final descriptor list is used to generate a fixed-length binary vector for each molecule.

The third type of descriptor (5P) is based on a fixed catalog of five simple molecular properties: hydrogen-bond acceptor counts, hydrogen-bond donor counts, molecular weight, number of rotatable bonds, and number of rings (using an SSSR definition). Note that this descriptor set utilizes exact counts rather than converted binary values.

For each of these three descriptor types, the final set used for model building is derived using a one-time iteration over the entire set of training data. This ensures that the identical descriptors are used in both the class-based and default approaches. In this way, any differences in predictive performance can be attributed to differences in the training protocol rather than the descriptor sets.

Training of Recursive Partitioning Models. Recursive partitioning refers to a particular class of decision-tree learning method based on the successive division of data using a single descriptor at a time. Each iteration of the method involves partitioning the current set of data into two disjoint groups, one for which the value of a given descriptor is less than a threshold, and the other for which the value is greater. The choice of the particular descriptor and threshold is based on the goal of producing partitions that are pure with respect to activity, such that active samples will tend to be placed in one partition and inactive samples in the other. The RP method applies this procedure to each new partition in an iterative fashion until some stopping criterion is achieved, ordinarily when each terminal partition has a high degree of purity. The implementation of RP used in the ChemTK application follows closely the description of Breiman et al.,[34] using the Gini metric to measure node purity, a minimum node size of two samples for default trees, a pruning procedure to control model complexity, and a misclassification metric—the fraction of molecules incorrectly classified as active or inactive—as the criterion to measure both training and testing performance.

Since the comparison of class-based and default approaches will be based on an external hold-out data set, it is important that the models be trained using a procedure designed to prevent overfitting. The goal for each approach is to identify an "optimal" model from the standpoint of future (rather than current) predictive performance, necessitating a resampling protocol to estimate the expected behavior of each model for unseen data. Based on this objective, and drawing from common approaches outlined in the learning-theory literature,[35] the following cross-validation and model-selection procedure was implemented:

(1) Initialize the model to have the highest level of complexity. For RP models, the highest complexity is associated with the default (full) decision tree, which in this study is grown until misclassification errors are minimized.

(2) Use k-fold cross-validation to train a two-class (active/inactive) classification model. The cross-validated classification performance is an unbiased estimate of future predictive ability for the specified level of model complexity. In our experiments we use $k = 4$.

(3) Incrementally decrease model complexity and repeat step (2). For RP models, complexity is decreased by an optimal pruning procedure, whereby the least predictive nodes are removed from the decision tree (see ref 34). This procedure is continued until the complexity has reached the desired lower limit, which for RP models corresponds simply to the single root node.

(4) Select the optimal complexity according to the model with best cross-validated performance.

(5) Construct a final model having the ideal complexity obtained in step (4) and using all of the training data (e.g.,

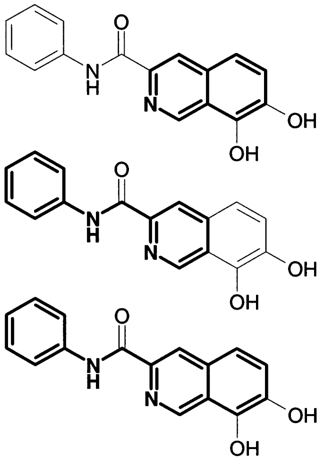

Figure 2. A sample NCI molecule in which every possible ring-system scaffold having between two and four rings is highlighted. Identification of scaffolds is as described in the text. Note that all connecting chains in a scaffold must be unbranched; accordingly, all highlighted atoms are of degree 2 or higher.

without cross-validation). This model represents the one best suited to perform predictions on future data and will be used on a hold-out set to compare class-based and default approaches. This comparison procedure is further explained in subsequent sections.

Class Generation Procedure. In this study, classes are generated on the basis of common scaffold, where each scaffold is composed of a single connected ring system. A connected ring system is defined as any number of single or fused rings joined together by an unbroken chain of atoms. Examples include, in increasing order of complexity, a single ring (benzene), a single fused system (naphthalene), and both the two previous systems (benzene and naphthalene) connected together by a chain of carbon atoms. Any connecting chains must be unbranched, such that all atoms in the final scaffold are connected to at least two other atoms.

The computational procedure used in ChemTK for identifying all such ring-system scaffolds from a molecular graph is as follows:

(1) Identify rings using a standard SSSR procedure. Here we use a protocol based on that of Figueras.[36]

(2) In an iterative fashion, eliminate from the graph all combinations of rings identified in step (1). Ring elimination involves the explicit deletion of each ring atom having a degree equal to 2 and all bonds connected to such an atom. Here the degree of an atom is defined by the number of neighbor atoms to which it is bonded. Following the deletion of each combination, proceed to step (3) only if the resulting graph remains a single connected system; if multiple fragments exists, consider the next combination.

(3) Using the graph obtained in step (2), iteratively eliminate all first-degree atoms that have been created. The resulting graph, which will now contain only atoms of degree 2 or higher, will define a scaffold that can be subsequently used for class generation.

Examples of ring-system scaffolds are shown in Figure 2. Note that in this study we have imposed the additional requirement that each scaffold contain between two and four rings. The lower limit in particular is imposed so that each class will be guaranteed a reasonable degree of structural homogeneity.

Classes are generated from an initial set of molecules using three steps. First, all possible ring-system scaffolds are identified from the molecules using the procedure outlined above and are stored in a single unique list. Next, each resulting scaffold is used to compile a list containing all molecules that have that scaffold as part of their molecular structures. Finally, scaffolds having a molecule list under a threshold size are eliminated. In the present study, we have chosen to keep only those scaffolds that represent at least 16 molecules; this is so that any predictive models subsequently built within a class will represent a sufficient number of training samples to be robust. The final set of scaffolds, and associated molecule lists, will now represent a complete scaffold-based organization of the initial data set.

There are a number of advantages to this class-generation method, particularly in the context of developing predictive models within each class as a subsequent process. The first is that, because they are each characterized by a specific scaffold, all classes have a precise chemical definition. This property, which ensures that any future (virtual) molecule can be reliably partitioned into the correct classes through a simple substructure search, means that each class-based QSAR model can be applied strictly to those molecules belonging to the same region of chemical space for which the model was trained. This concept of reliability is elaborated in a future section. The second advantage is that each class is guaranteed to have a reasonably high degree of structural homogeneity, such that any pair of molecules from a given class could, at least plausibly, operate by a similar biological mechanism. This property, also discussed at length in later sections, is particularly important for generating QSAR models with a high degree of human interpretability. Note that while several scaffold-based algorithms similar to, and perhaps superior to, this one have been described previously,[28,37,38] many popular methods, and in particular "fingerprint" methods based on small, discontinuous fragments, often fail with respect to both of the advantages just discussed, being neither able to assign future data easily, nor to yield classes that are structurally homogeneous in all cases.

Two potential drawbacks to our method should also be noted. The first is that the number of unclassified, or "singleton," molecules may be reasonably high, both because some of the initial molecules may be acyclic and thus unable to contain any ring-system scaffold, and because our 16-molecule requirement on class size may lead to the omission of some relatively rare classes having fewer members. The second drawback is that the final set of classes is likely to be redundant, with a given molecule falling into more than one chemical class as a result of having more than one ring-system scaffold. While this redundancy could be reduced by applying any standard variable-selection procedure,[39] we have decided to permit the redundancy to remain, so that no additional bias is introduced by the arbitrary elimination of scaffolds. Each of these potential drawbacks is further discussed in later sections.

Experimental Protocol for Method Comparison. The comparison of class-based and default approaches is based on the use of five random trials. In each trial an initial data set is assembled using the full collection of 1501 active molecules and a randomly chosen collection of 5000 inactive molecules (from the initial 41 049), yielding a total set of

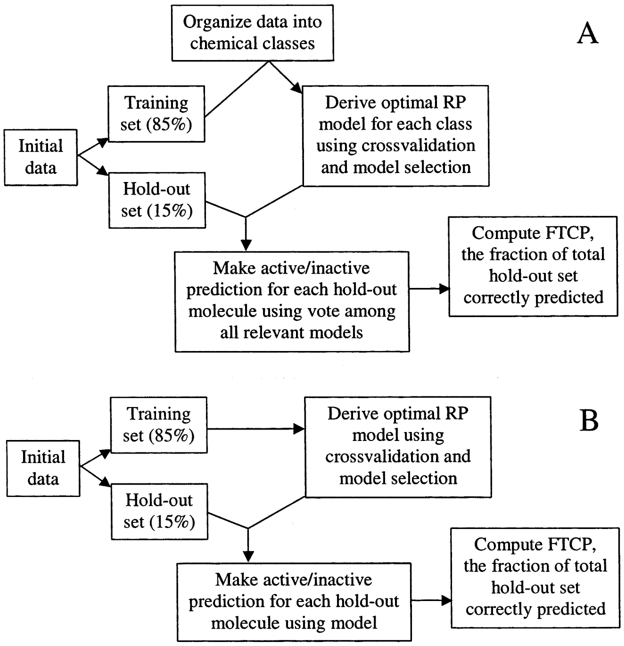
A
Organize data into chemical classes
Training set (85%)
Initial data
Hold-out set (15%)
Derive optimal RP model for each class using crossvalidation and model selection
Make active/inactive prediction for each hold-out molecule using vote among all relevant models
Compute FTCP, the fraction of total hold-out set correctly predicted
B
Training set (85%)
Initial data
Hold-out set (15%)
Derive optimal RP model using crossvalidation and model selection
Make active/inactive prediction for each hold-out molecule using model
Compute FTCP, the fraction of total hold-out set correctly predicted

Figure 3. Protocols for model construction. The class-based protocol (A) requires that training data first be organized into classes, and that a separate model be constructed for each resulting class. Screening is then performed using a consensus strategy, as explained in the text. The default protocol (B) involves only a single model constructed using all of the training data at one time. Both protocols are applied to five separate trials, each using a different randomly selected initial data set.

size 6501, of which only the active portion is identical across each trial. We follow this approach of selecting only a fraction (12%) of the inactive data in order to reduce the time required for model training and to prevent a severely skewed active/inactive ratio that would require special modeling techniques; subsequent control tests of larger inactive subsets (up to 25%) produced nearly identical results. Each of these five sets is then further divided into a training portion consisting of 85% of the data (5525 molecules) and a testing, or "hold-out," portion consisting of 15% (976 molecules). The latter set is used to provide an unbiased comparison of the two model-generation approaches, as described below. In each trial this 85/15 division is random, but it is ensured that the two subsets contain identical proportions of biological result values (CA, CM, and CI).

Once the five data sets are constructed, each is used in two separate experiments, one representing a **class-based** approach to model generation and the other a **default** approach. The class-based approach is summarized in Figure 3A. Here the training data are first subjected to a class-generation process that organizes the data into chemical classes defined by ring-system scaffolds, as described in an earlier section. Note that the resulting set of classes can be redundant, such that a given molecule can belong to more than one class. Once classes are generated, each is used to generate an optimal RP model using the cross-validation and model-selection procedure outlined earlier. This model-generation step is performed several times, once using each of the distinct descriptor types (pharmacophore, fragment, and property-based) described previously.

Each set of class-based models—one set for each descriptor type—is then used to test the hold-out data. In this testing procedure, which must take into account multiple models, each hold-out molecule is first assigned to all relevant chemical classes identified in the initial training step. Assignment to a given class occurs only if the hold-out molecule contains the ring-system scaffold used to define that class; a given hold-out molecule can therefore belong to any number of classes, including zero. For those hold-out molecules belonging to at least one class, separate active/inactive predictions are made using all of the relevant class-based RP models, and a single "consensus" prediction is made by taking a vote among all of these individual class-based predictions (with ties arbitrarily resulting in a prediction of active). Thus each hold-out molecule receives only a single activity prediction regardless of the number of chemical classes to which it belongs. The final predictive performance of the class-based procedure is then based on the fraction of the hold-out set correctly predicted as being either active or inactive. Because this performance metric is derived using a hold-out set never seen during the training procedure, it can be used in an unbiased manner to compare the two model-generation approaches. Note that molecules not belonging to any ring-system class receive no active/inactive predictions and accordingly do not contribute to the class-based performance metric. The effect of these molecules is therefore to reduce the "coverage" of class-based models; the Results section quantifies this effect and provides further discussion.

Thus in summary, the class-based approach to model generation involves the following progression of steps:

For each random trial (5 total):

For each type of molecular descriptor (3 total):

(1) Generate class-based models using training data

(2) Test class-based models on hold-out data using consensus scoring, and record performance.

The default approach to model generation is illustrated in Figure 3B. In this protocol, the training data are no longer subjected to the initial class-generation procedure described previously. Instead, a single optimal RP model is constructed from all of the training data at one time, using the identical cross-validation and model-selection protocol as in the class-based approach. Once a model is constructed, it is used as before to test the hold-out portion of the data. In this case each hold-out molecule receives only a single active/inactive prediction, and the overall predictive performance is measured as the fraction of the hold-out set correctly predicted as being either active or inactive.

As before, model generation is performed separately for each of several descriptor types, but in the default approach two new types are added to the three used previously. The first is based on the same list of ring-system scaffolds identified during the class-generation process of the class-based approach. This list of structures, used to generate a fixed-length binary vector via substructure search, is intended to serve as a control to determine whether the scaffolds used in class generation might themselves be predictive of activity. The second set of descriptors is simply a concatenation of the ring-system descriptors and the unbranched-fragment descriptors. These two additional descriptor sets are referred to by the abbreviations "RS" and "UF+RS."

One significant addition is required to complete the training and testing protocols for the default case. Because the class-based approach requires that hold-out molecules first be tested against a set of structural classes, it is possible that a significant number of these molecules will be

Table 1. Results of Class Generation for Five Random Trials

trial	sample size	number of classes	molecules classified	average class size	redundancy
1	5525	147	2690	39.7	2.2
2	5525	145	2649	39.8	2.2
3	5525	145	2700	40.1	2.2
4	5525	136	2654	41.2	2.1
5	5525	143	2695	41.3	2.2

designated as singletons and will thus never contribute to the final hold-out performance value. In contrast, the default approach guarantees a prediction for each hold-out molecule, so the resulting performance value will always reflect the entire set of hold-out data. As a result, it may be difficult to make a direct comparison between the performance values of these two approaches. To address this problem, a supplementary protocol is used for default models. In addition to building models using the full 5525-molecule training set, models are also built using only the subset of these data successfully classified in the corresponding class-based approach. In other words, for a given random trial, all training molecules not designated singletons in the class-based approach are used to train a separate RP model in the default approach. This secondary model is then tested against the analogous "classified" subset of the hold-out data set (e.g., only those hold-out molecules not designated singletons in the hold-out phase of the class-based approach). Thus there are two separate default approaches, referred to as **full** and **reduced**, respectively, and each is used to generate a separate hold-out performance value. Since it is not clear which of the two values is most appropriate for comparison with the class-based counterpart, both will be presented in the subsequent Results section. Note that a third default approach, using the full 85% set for training and the reduced (classified) portion of the hold-out set for testing, was also pursued. This third approach gave the poorest overall results and is not presented here.

Thus in summary, the default approach to model generation involves the following progression of steps:

For each random trial (5 total):

For each training set (2 total):

For each type of molecular descriptor (5 total):

(1) Generate model using training data

(2) Test model on appropriate hold-out data (full or reduced), and record performance.

RESULTS

Table 1 summarizes the results of class generation for the five random trials. The table shows, for each trial, the number of molecules used in the classification, the number of classes generated, the number of molecules successfully classified (i.e., the number that do not become singletons), the average number of molecules per class, and the class redundancy (the number of classes, on average, in which each molecule appears). Note that most of these values differ across trials, due to the use of a different random subset of the initial data in each instance.

The average number of classes generated in each trial is 143, and the average class size is 40 molecules, with sizes ranging from 16 molecules (the required minimum) to a maximum of 289 (occurring in trial 5). The significant classification redundancy, at 2.2 classes per molecule on average, results from the use of each identified scaffold to create a distinct class; while such redundancy could easily be reduced by pruning the initial scaffold list, we felt it best to eliminate any bias that such a procedure might introduce. Meanwhile, the classification coverage is somewhat low, with only around 49% of training molecules on average falling into at least one class. This result is due both to our classification method, which automatically rejects all acyclic molecules, and to our relatively high limit for the minimum class size. While this effect could likewise be addressed in the future using only modest alterations to the protocol, it does nonetheless pose a potential limitation to the class-based approach, since a similarly low coverage can be expected during the hold-out phase of the procedure (see below). Further discussion of the low coverage value will be provided in subsequent sections.

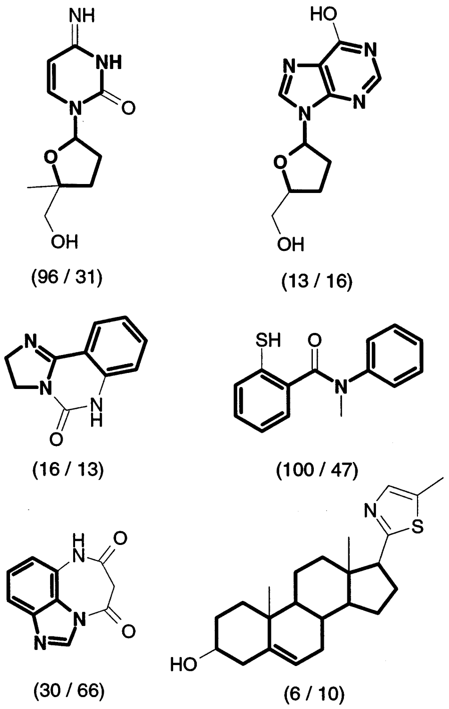

Figure 4. Sample ring-system classes. Each molecule represents a single scaffold-based class identified from the NCI training data, with corresponding scaffolds shown highlighted (examples are from trial 1). The two values beneath each molecule are the numbers of training actives and inactives within the corresponding class.

Figure 4 shows six sample classes from the set of 147 generated in trial 1. In each case, the defining ring-system scaffold is shown highlighted, and the accompanying molecule represents the class member having the lowest molecular weight. Note that some of the scaffolds shown in the figure represent only one of several equally valid possibilities for the accompanying molecule; this scaffold degeneracy leads to the classification redundancy discussed earlier. Not unexpectedly, each of the scaffolds in the figure is also identified in each of the four remaining random trials, a result that reflects the high degree of overall similarity between the five classifications. This similarity is likely due to both the large sizes of the random data sets and the requirement of a large minimum class size.

Descriptors were derived for the training molecules of each random trial as previously described. Recall that since descriptor generation is based on a one-time processing of

Table 2. Results of Class-Based and Default Model Generation

row	method	descriptors[a]	tree size	hold-out coverage	FACP[b]	FICP[c]	FTCP[d]
1	class based	UF	2.6	0.46	0.70	0.89	0.83
2	class based	5P	2.1	0.46	0.72	0.85	0.81
3	class based	PH	2.4	0.46	0.71	0.84	0.80
4	class based	none	1.0	0.46	0.65	0.83	0.77
5	full default	UF	41.0	1.00	0.38	0.96	0.83
6	full default	5P	3.4	1.00	0.23	0.95	0.78
7	full default	PH	33.4	1.00	0.16	0.98	0.79
8	full default	RS	22.6	1.00	0.34	0.97	0.82
9	full default	UF+RS	40.2	1.00	0.40	0.96	0.83
10	reduced default	UF	35.4	0.46	0.63	0.89	0.81
11	reduced default	5P	9.0	0.46	0.41	0.90	0.75
12	reduced default	PH	27.8	0.46	0.31	0.95	0.75
13	reduced default	RS	27.0	0.46	0.58	0.90	0.80
14	reduced default	UF+RS	32.6	0.46	0.62	0.89	0.81

[a] Unbranched fragments (UF), fixed properties (5P), pharmacophores (PH), ring systems (RS), or some combination thereof. [b] The fraction of active molecules correctly predicted. [c] The fraction of inactive molecules correctly predicted. [d] The fraction of total molecules correctly predicted.

each of the five training sets, identical final descriptor sets are used for both the class-based and default protocols (though the sets vary slightly across random trials). The number of descriptors of each type, averaged over the five trials, are 2354 (UF), 5 (5P), 225 (PH), 143 (RS), and 2497 (UF+RS).

The comparison of class-based and default approaches to model generation is summarized in Table 2. Each row shows the results, averaged over five trials, of a single experiment described in the earlier protocol section and is characterized by both a training method (class-based, full default, or reduced default) and a descriptor set (UF, 5P, PH, RS, or UF+RS), listed in the second and third table columns, respectively. The five remaining columns provide performance characteristics for each specified method/descriptor combination. Column 4 shows the average number of nodes in the associated RP trees, while column 5 shows hold-out coverage, defined as the fraction of all hold-out molecules receiving a prediction from the derived models. Note that only in the full-default approach does this coverage value represent the complete hold-out set. In the class-based approach—and therefore also the reduced-default approach, since it uses the identical hold-out data—only 46% of hold-out molecules are successfully classified by any ring-system scaffold, so only this portion can be tested using the associated RP models (more discussion of coverage is provided below). The final three columns describe the average quantitative performances of the RP models in each experiment. **FACP** is the fraction of active hold-out molecules correctly predicted as active by the model. **FICP** provides the analogous metric for inactive hold-out molecules. **FTCP** is the fraction of the total hold-out set correctly predicted as either active or inactive. It is this final value that provides the principal basis for comparing results of the three training approaches.

Using the FTCP metric for comparison, the table indicates that the unbranched-fragment (UF) descriptors, or some combination thereof, provide the best results in each of the three model protocols, with corresponding values of 83% (class-based), 83% (full default), and 81% (reduced default). These values are consistent with those found in other studies using the same NCI data.[11,27] The other descriptor types, most likely due to the much smaller sizes of their sets (see above), are uniformly inferior to their UF-based counterparts, yielding values of 81% (5P) and 80% (PH) for the class-based approach, 78% and 79% for the full-default approach, and 75% and 75% for the reduced-default approach. Since most of these values, and in particular those based on the UF descriptors, do not represent differences that are statistically significant (derivations not shown), the preliminary conclusion is that all three approaches are reasonably similar in terms of predictive performance.

In exploring the comparisons further, we find that of the two default approaches (full and reduced), it is the latter that likely provides the most relevant comparison to the class-based approach, even though it is the former that yields the best overall performance values. The reason is that while the class-based and reduced-default approaches utilize identical data (both for training and testing), the full-default data, which differ in composition due to the higher model coverage rates for that approach (column 5 of Table 2), have a substantially different active-to-inactive ratio than their reduced-default counterpart: while the fraction of active molecules in the reduced hold-out set is 32% on average, in the full set it is only 23%. This difference has a large impact on the expected performance values for the associated models, a point illustrated by the following argument: simply by predicting that every hold-out molecule is inactive, the full-default model can achieve an FTCP value of 77%, while the reduced-default can achieve only 68%. Thus it is likely that the performance values obtained for the full-default method are artificially inflated and that only the reduced-default values should be used in the comparisons. The much closer similarity between both FACP and FICP values for the class-based and reduced-default approaches supports this conclusion: while FACP values of over 60% are achieved in both of these approaches, the full-default approach yields no FACP value larger than 40%.

Given the greater apparent reliability of the reduced-default versus the full-default results, the class-based approach appears even more favorable. Since the class-based FTCP values from Table 2 are larger than the reduced-default values by statistically significant margins for all but the UF descriptors, we conclude that with respect to predictive performance, the class-based approach is at least equal to the default and may in fact be superior, particularly in the context of allowing greater flexibility in the choice of descriptor type.

It is somewhat surprising that the class-based approach would perform so well, given the inherent inflexibility of its training protocol. In this approach, the initial class-generation step uses a so-called “unsupervised” method: that is, one designed strictly for the purpose of organizing data by similarity and not for the purpose of predicting activity values. Since this step puts a significant constraint on any subsequent data analysis, one might assume that the predictive performance of any class-based models would be adversely affected relative to those of a default approach in which the constraint is absent. In other words, an entirely “supervised” approach designed for activity prediction, and in particular one of high flexibility such as our default approach combining RP and the large UF or UF+RS descriptor sets, should outperform any similar protocol that incorporates a significant unsupervised routine.

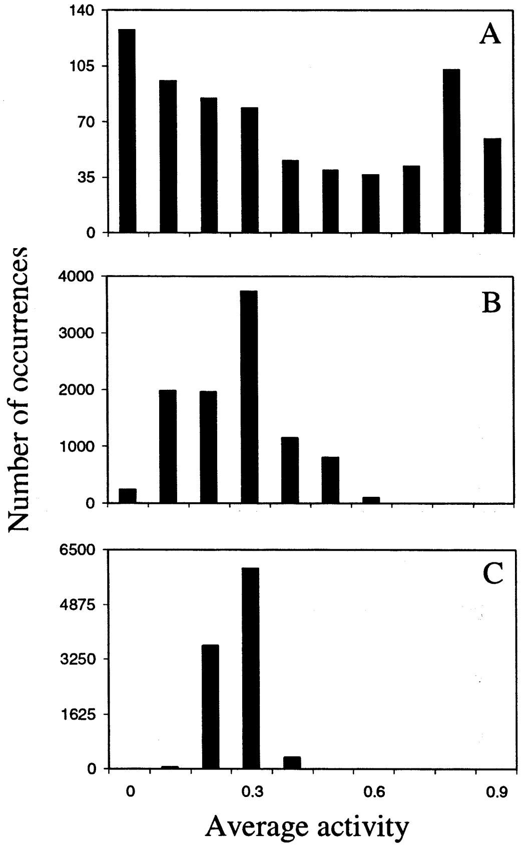

Figure 5. Average activity distributions for classes. The *x*-axes represent the average activity of molecules in a class, using 0.0 for inactives and 1.0 for actives. The *y*-axes represent the number of classes in a given activity bin. The top diagram (A) shows the distribution for all ring-system classes identified in the class-based approach. The lower two diagrams show corresponding control distributions, where classes of size 16 (B) or 100 (C) are chosen randomly from the identical data set as in part A. The bimodal shape of the upper graph suggests the tendency for each scaffold-based class to group together molecules of similar activity.

Figure 5 may help to explain this result. The upper diagram (5A) shows a distribution of average activity values from all of the classes generated using the class-based protocol (716 classes from all five trials). For comparison, the two lower diagrams (5B and 5C) show analogous control distributions that result when 10 000 “fictitious” classes, of size 16 and 100, respectively, are chosen at random from the identical pool of training data that is used in Figure 5A. In all three cases activity values are treated as either 0 (inactive) or 1 (active). The figure shows that whereas the latter two plots are, as expected, distributed uniformly around the true population average of 0.31, the plot in 5A is significantly bimodal, with apparent peaks near activity values of 0 and 1. This behavior suggests that our class-generation procedure by itself tends to group molecules on the basis of activity, producing many classes with average activities significantly higher than, or significantly lower than, the expected population average.

To confirm this result, we repeated the class-based experiments described earlier, but this time using an empty descriptor set. The resulting RP models therefore consisted of only the single root node, such that subsequent predictions were based only on the ratio of active to inactive molecules within the corresponding classes. The resulting FTCP value of 77%, shown in row 4 of Table 2, is nearly as high as those produced with the nonempty descriptor sets in the class-based approach and is actually higher than several of the values obtained using the reduced-default approach. Though the reason for this behavior is not entirely clear, it likely stems from the well-known and frequently cited idea in computational and medicinal chemistry that molecules of similar structure tend to show similar patterns of activity. This principle, often cited as the rationale for “similarity search” methods,[40−42] would support our finding that a straightforward class-generation method, particularly one based on common scaffolds where each class is assured a reasonably high degree of structural homogeneity, can by itself demonstrate good performance as a predictive tool. More thorough study is required to determine whether this result is mainly an artifact of the NCI data set, which clearly contains a large number of distinct, highly populated chemical families.

DISCUSSION

Although the difference in quantitative performance between the class-based and default approaches is relatively small, there are two areas in which the class-based approach presents a clear advantage in our study. The first is interpretability, as measured by the ease with which useful SAR information can be extracted from the resulting models. The second is the reliability with which such models can be applied to new data in a virtual-screening application. Each of these topics is discussed below.

As Table 2 indicates, the class-based approach produces models that are considerably less complex than the default models, as measured by the numbers of nodes contained in the final RP trees. This lower complexity, along with the significant degree of structural homogeneity imparted to the ring-system classes by their representative scaffolds, makes SAR extraction from class-based models particularly straightforward. Figure 6 illustrates three such examples, each representing a complete RP model from the class-based approach. The molecules on the left represent individual classes, the defining scaffolds of which are shown highlighted. Each adjacent chart represents the two root-node partitions of the corresponding class-based RP tree: the two pairs of bars on the left and right indicate the activity effects of the absence and presence, respectively, of the particular structural element circled on the representative molecule. From each of these examples it is possible to extract a simple and specific statement describing the SAR. For instance, one can infer from the first example (6A) that “in the context of benzimidazole derivatives, the presence of a thiazolidine ring fusion is activating”. In this case the relevant descriptor is of the UF type (Smarts pattern “CSCN”), and the activation effect is indicated by the pair of bars at the far right: of the 21 benzimidazoles containing this particular substitution pattern, 17 are active and only four are inactive. Similar interpretations are possible for Figure 6B,C, the former revealing an activating pharmacophore in the context of steroids, and the latter a deactivating fragment in the context of purine nucleosides. In all of these cases, the simplicity and specificity of the inferred SAR are made possible by the low model complexity and high structural homogeneity inherent in the class-based approach. Because each model requires only a small number of relevant descriptors, and because each common scaffold provides a chemically

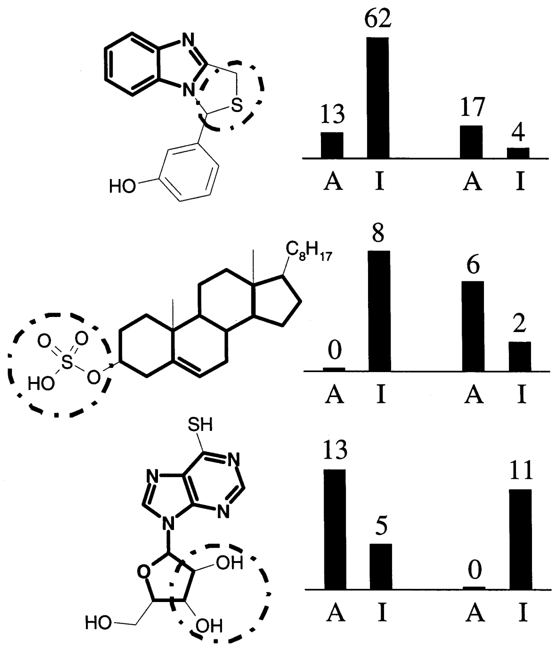

Figure 6. Examples of SAR from class-based models. Each molecule represents a distinct ring-system class as defined by the highlighted scaffold. Dashed circles indicate the molecule region where the relevant SAR is observed. Charts at the right, each derived from a complete single-split (three-node) RP tree, indicate activity effects: the two pairs of bars at the left and right show the numbers of training actives (A) and inactives (I) that lack and contain, respectively, the descriptor element from the model. These descriptors are, respectively, the four-atom fragment (Smarts "CSCN") representing the fused thiazolidine ring; a graph-based pharmacophore of three hydrogen-bond acceptors, each pair separated by two bonds; and the four-atom fragment (Smarts "OCCO") representing a 2′,3′ disubstitution of any two −OR groups. All examples are from the first random trial.

relevant context, the class-based approach yields interpretable activity patterns that, even if revealing only correlative relationships and not true causal effects, can nonetheless have practical utility in both the formulation of new hypotheses and the design of future experiments.

In contrast, the significantly larger default models do not lend themselves to as straightforward an SAR analysis. As an illustrative example we have analyzed in detail the RP tree constructed in the first trial of the reduced-default approach using the UF+RS descriptor set (row 14 of Table 2). A simple structural representation of this 37-node tree is shown in Figure 7. It was expected that this particular tree, despite its size, might be among the most highly interpretable default models, due to its smaller training set and incorporation of the large ring-system structures as descriptors. In fact, it was thought that this tree might largely reproduce the structural classification and SAR results observed in the class-based analysis, with the initial RP splits dividing the data by broad chemical family, and subsequent splits within those families providing easy-to-interpret SAR similar to that of class-based trees. Instead, we find that most chemical families are never isolated within individual nodes but rather remain intermingled in highly heterogeneous groups. For example, of the three structural classes illustrated in Figure 6, 70% of the associated molecules—59 of 96 benzimidazoles, 13 of 16 steroids, and 26 of 29 nucleosides—fall into the identical terminal node of the tree (shown highlighted in Figure 7). In fact, this "catch-all" node, which is typical of our default RP trees and of the RP method generally, contains 63% of the entire training set and includes representatives from nearly every distinct structural class present in the data. In the context of such a heterogeneous group of molecules, any simple and specific statement describing the SAR, such as the benzimidazole ring fusion example cited earlier, is extremely difficult to make, because there is no identifiable structural context to help define it. Rather, this example suggests that given structurally diverse initial data, the RP method, and likely any other supervised method driven only by structure−activity correlations and ignoring structural homogeneity, is unlikely to yield models that are as easily understood and applied, even if overall predictive performance is high (81% FTCP in this example). Note that the difficulty in interpreting RP models, and in particular the failure of initial splits to partition by broad chemical family, has been discussed by other authors, many of whom have proposed modifications that address this problem directly. These efforts are discussed in more detail below.

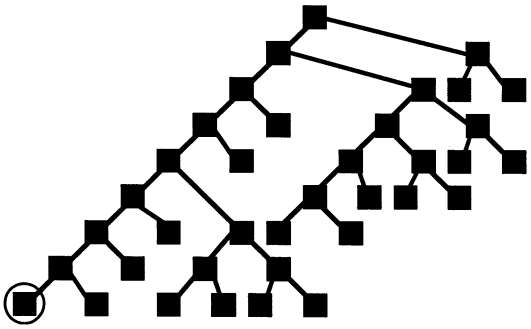

Figure 7. Representation of the RP tree constructed in trial 1 using the reduced-default approach and the UF+RS descriptor set. The circled node at the lower left contains 63% of the training data and is highly heterogeneous as a consequence.

A second advantage of the class-based approach is the reliability with which resulting models can be used in virtual screening applications. The reason has to do with "interpolation" versus "extrapolation," respectively the application of predictive models to data similar to, versus different from, the training data. Because each class-based model is associated with a specific scaffold, the data used in training that model are not only homogeneous but also precisely defined in chemical terms. As a consequence, a subsequent virtual screening protocol can incorporate a substructure search to reject automatically any molecules not fitting into this precise definition, thereby accepting only those molecules of the appropriate chemical family. While such an approach does lead to lower coverage, as we have seen, it has the benefit of avoiding extrapolation into regions of chemical space not appropriate for the particular QSAR model. In contrast, with general models constructed using diverse data, there is usually no corresponding definition for the relevant chemical space of the training set, so it is difficult to avoid such extrapolation. This leads to screening results that are suspect, since the model might not be adequately trained in the entire chemical region spanned by the virtual library.

There are also some potential disadvantages to the class-based approach. First, as we observed in the present work, a significant portion of training and test molecules may be dropped from the analysis, since they will not belong to any particular class. Second, the number of class-based models can become quite large (143 class-based models in our study, on average, versus a single default model), potentially eliminating any advantage gained by their relative simplicity and ease of interpretation. Third, since class-based models are derived within well-defined chemical families, they may provide little help in the important task of identifying new lead candidates of novel structural origin.

Each of these seeming disadvantages can, however, be largely overcome. First, model coverage could likely be increased by improving the class-generation procedure. The protocol used in these experiments, while useful for illustrating the general class-based approach, produces an unusually large number of singletons; a reduced emphasis on ring systems, and a lower threshold on class size, could improve this result. Second, while the number of class-based models may be large, the fact that they are automatically partitioned by chemical family greatly facilitates their analysis. In particular, given our ring-system approach to classification, each of the resulting models is characterized by a well-defined chemical scaffold, making it relatively easy to organize, browse, and prioritize the complete set of models using simple visual inspection. Furthermore, by lowering the classification redundancy, such as by pruning the initial scaffold list, one could significantly reduce the number of models from what was obtained in the current study. Third, although class-based models are developed within well-defined chemical families, it may be possible to combine several such models together in order to produce a more "general" model; such a combined model may have applicability to structural classes outside of any used during the training phase. This last issue is discussed below as a future direction.

In summary, our study has shown that a class-based approach to model generation can demonstrate equal predictive performance to a similar default approach in which no organizational step is incorporated. More significantly, this approach shows a clear advantage in producing models that are more easily understood and applied in the context of a chemical design project involving diverse initial data. As an interesting comparison to these results, a number of novel strategies have been recently devised to increase model interpretability more directly, using modifications to standard algorithms such as recursive partitioning.[29,43] These efforts not only represent important advances in QSAR research but also are likely the preferred approach from the standpoint of pure predictive accuracy, since they more closely integrate the organizational and predictive components into a single method. In contrast, our class-based approach illustrates how these two processes can be completely separated, provided that there is a formal mechanism by which they can be tested together as a single procedure, as in our consensus-scoring and hold-out experiments. A potential advantage is that the organizational step can now be designed independently for the express purpose of increasing the interpretability of the final class-based models and can borrow from any number of widely available and well-described algorithms. In combining this step with an equally general learning method (RP or otherwise), the resulting class-based models might, if our experiments are sufficiently general, suffer very little cost in terms of predictive performance relative to models built using the same learning method on its own. The relative simplicity of this approach makes it an interesting alternative, and a useful complement, to the development of more advanced learning algorithms.

Future directions of research will be focused within two main areas. First, we plan to repeat our experiments using a wider range of descriptor types, learning algorithms, and class-generation methods. Since the conclusions of our study are intended to be general ones pertaining to a broad class-based strategy and not to any particular set of protocols or techniques, it is important that our methods be expanded beyond the relatively simple choices used for the initial tests. Second, we plan to explore the development of more general, class-independent models using the class-based models as "building blocks." A considerable amount of recent work has focused on the development of such general models, typically for the purpose of "evolving" from one chemical class to another.[44] This lead-evolution process can clearly be a desirable strategy in drug discovery, particularly where a lead compound may present toxicological or other problems, and requires models that can describe chemical space in a relatively broad sense. A danger with deriving such models directly from diverse data is the potential to uncover spurious correlations, since the vast size of chemical space virtually guarantees that two structurally dissimilar molecules with no shared biological mechanism can nonetheless have nearly identical abstract (e.g., pharmacophore) representations. Here the class-based approach can have clear advantages, in that the initial within-class models can first be validated individually on smaller, focused data sets with sufficient homogeneity to provide confidence in performance. By then explicitly testing, in a statistically controlled manner, whether such models can be predictive within other classes, one might be able to identify models with applicability to a wider range of chemical families, while at the same time reducing the risk of drawing erroneous or misleading conclusions. In a preliminary study of this idea, we have made a pairwise comparison of all class-based RP models of a fixed descriptor type and trial number and have identified a significant number that are identical or highly similar, an indication that some SAR information is likely to be independent of any particular chemical scaffold. Subsequent work will test whether these "equivalent" class-based models, or some type of "average" model obtained by combining them, can perform true active/inactive discrimination within other classes, and if so, whether their performance surpasses that of default models constructed directly from the same data.

REFERENCES AND NOTES

(1) Hansch, C.; Muir, R. M.; Fujita, T.; Maloney, P. P.; Geiger, E.; Streich, M. The Correlation of Biological Activity of Plant Growth Regulators and Chloromycetin Derivatives with Hammett Constants and Partition Coefficients. *J. Am. Chem. Soc.* **1963**, *85*, 2817−2824.

(2) Fujita, T.; Iwasa, J.; Hansch, C. A. New Substituent Constant, π, Derived From Partition Coefficients. *J. Am. Chem. Soc.* **1964**, *86*, 5175−5180.

(3) Burden, F. R. Using Artificial Neural Networks to Predict Biological Activity from Simple Molecular Structural Considerations. *Quant. Struct.-Act. Relat.* **1996**, *15*, 7−11.

(4) King, R. D.; Hirst, J. D.; Sternberg, M. J. E. New Approaches to QSAR: Neural Networks and Machine Learning. *Perspect. Drug Discov. Des.* **1993**, *1*, 279−290.

(5) Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design*; Wiley & Sons: New York, 1999.

(6) Greene, N.; Judson, P. N.; Langowski, J. J.; Marchant, C. A. Knowledge-Based Expert Systems for Toxicity and Metabolism

Prediction: DEREK, StAR, and METEOR. *SAR QSAR Environ. Res.* **1999**, *10*, 299−314.

(7) Enslein, K.; Gombar, V. K.; Blake, B. W. Use of SAR in Computer-Assisted Prediction of Carcinogenicity and Mutagenicity of Chemicals by the TOPKAT Program. *Mutat. Res.* **1994**, *305*, 47−61.

(8) Klopman, G. Artificial Intelligence Approach to Structure−Activity Studies. Computer Automated Structure Evaluation of Biological Activity of Organic Molecules. *J. Am. Chem. Soc.* **1984**, *106*, 7315−7321.

(9) Rusinko, A., III.; 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.

(10) Chen, X.; Rusinko, A., III.; Young, S. S. Recursive Partitioning Analysis of a Large Structure−Activity Data Set Using Three-Dimensional Descriptors. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1054−1062.

(11) Miller, D. W. Results of a New Classification Algorithm Combining K Nearest Neighbors and Recursive Partitioning. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 168−175.

(12) Stanton, D. T.; Morris, T. W.; Roychoudhury, S.; Parker, C. N. Application of Nearest-Neighbor and Cluster Analyses in Pharmaceutical Lead Discovery. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 21−27.

(13) Hoffman, B.; Cho, S. J.; Zheng, W.; Wyrick, S.; Nichols, D. E.; Mailman, R. B.; Tropsha, A. Quantitative Structure−Activity Relationship Modeling of Dopamine D_1 Antagonists Using Comparative Molecular Field Analysis, Genetic Algorithms-Partial Least-Squares, and K Nearest Neighbor Methods. *J. Med. Chem.* **1999**, *42*, 3217−3226.

(14) Goodwin, J. T.; Mao, B.; Vidmar, T. J.; Conradi, R. A.; Burton, P. S. Strategies Toward Predicting Peptide Cellular Permeability From Computed Molecular Descriptors. *J. Peptide Res.* **1999**, *53*, 355−369.

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

(16) Palm, K.; Luthman, K.; Ungell, A.; Strandlund, G.; Beigi, F.; Lundahl, P.; Artursson, P. Evaluation of Dynamic Polar Molecular Surface Area as Predictor of Drug Absorption: Comparison with Other Computational and Experimental Predictors. *J. Med. Chem.* **1998**, *41*, 5382−5392.

(17) Klopman, G.; Frierson, M. R.; Rosenkranz, H. S. The Structural Basis of the Mutagenicity of Chemicals in *Salmonella Typhimurium*: The Gene-Tox Data Base. *Mutat. Res.* **1990**, *228*, 1−50.

(18) Ajay; Walters, W. P.; Murcko, M. A. Can We Learn To Distinguish Between "Drug-Like" and "Nondrug-Like" Molecules? *J. Med. Chem.* **1998**, *41*, 3314−3324.

(19) Sadowski, J.; Kubinyi, H. A Scoring Scheme for Discriminating Between Drugs and Nondrugs. *J. Med. Chem.* **1998**, *41*, 3325−3329.

(20) Dearden, J. C.; Barratt, M. D.; Benigni, R.; Bristol, D. W.; Combes, R. D.; Cronin, M. T. D.; Judson, P. N.; Payne, M. P.; Richard, A. M.; Tichy, M.; Worth, A. P.; Yourick, J. J. The Development and Validation of Expert Systems for Predicting Toxicity. The Report and Recommendations of ECVAM Workshop 24. http://www.jhsph.edu/∼altweb/science/pubs/ECVAM/ecvam24.htm.

(21) Kubinyi, H. *QSAR: Hansch Analysis and Related Approaches*; VCH Publishers: New York, 1993.

(22) Devillers, J. *Comparative QSAR*; Taylor & Francis; Washington, DC, 1998.

(23) Bajorath, J. Selected Concepts and Investigations in Compound Classification, Molecular Descriptor Analysis, and Virtual Screening. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 233−245.

(24) Cramer, R. D., III.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.

(25) 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.

(26) Zheng, W.; Tropsha, A. Novel Variable Selection Quantitative Structure−Property Relationship Approach Based on the k-Nearest-Neighbor Principle. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 185−194.

(27) Klopman, G.; Tu, M. Diversity Analysis of 14156 Molecules Tested by the National Cancer Institute for Anti-HIV Activity Using the Quantitative Structure−Activity Relational Expert System MCASE. *J. Med. Chem.* **1999**, *42*, 992−998.

(28) Roberts, G.; Myatt, G. J.; Johnson, W. P.; Cross, K. P.; Blower, P. E. LeadScope: Software for Exploring Large Sets of Screening Data. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1302−1314.

(29) 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.

(30) For information regarding the ChemTK application, contact the author at dmiller@sageinformatics.com, or visit the Sage Informatics website at http://www.sageinformatics.com.

(31) National Cancer Institute. Bethesda, MD, http://www.nci.nih.gov.

(32) Livingstone, D. J. The Characterization of Chemical Structures Using Molecular Properties. A Survey. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 195−209.

(33) Greene, J.; Kahn, S.; Savoj, H.; Sprague, P.; Teig, S. Chemical Function Queries for 3D Database Search. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1297−1308.

(34) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees*; Wadsworth: Monterey, CA, 1984.

(35) Cherkassky, V. S.; Mulier, F. *Learning From Data. Concepts, Theory, and Methods*; Wiley & Sons: New York, 1998.

(36) Figueras, J. Ring Perception Using Breadth-First Search. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 986−991.

(37) Sheridan, R. P.; Miller, M. D. A Method for Visualizing Recurrent Topological Substructures in Sets of Active Molecules. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 915−924.

(38) Nicolaou, C. A.; Tamura, S. Y.; Kelley, B. P.; Bassett, S. I.; Nutt, R. F. Analysis of Large Screening Data Sets via Adaptively Grown Phylogenetic-Like Trees. *J. Chem. Inf. Comput. Sci*. **2002**, in press.

(39) Blum, A. L.; Langley, P. Selection of Relevant Features and Examples in Machine Learning. *Artificial Intelligence* **1997**, *97*, 245−271.

(40) *Concepts and Applications of Molecular Similarity*; Johnson, M., Maggiora, G. M., Eds.; John Wiley & Sons: New York, 1990.

(41) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.

(42) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do Structurally Similar Molecules Have Similar Biological Activity? *J. Med. Chem.* **2002**, *45*, 4350−4358.

(43) Cho, S. J.; Shen, C. F.; Hermsmeier, M. A. Binary Formal Inference-Based Recursive Modeling Using Multiple Atom and Physiochemical Property Class Pair and Torsion Descriptors as Decision Criteria. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 668−680.

(44) Srinivasan, J.; Castellino, A.; Bradley, E. K.; Eksterowicz, J. E.; Grootenhuis, P. D. J.; Putta, S.; Stanton, R. V. Evaluation of a Novel Shape-Based Computational Filter for Lead Evolution: Application to Thrombin Inhibitors. *J. Med. Chem.* **2002**, *45*, 2494−2500.

CI025606G