On Combining Recursive Partitioning and Simulated Annealing To Detect Groups of Biologically Active Compounds

Paul Blower,*,† Michael Fligner,‡ Joseph Verducci,‡ and Jeffrey Bjoraker†,§

Leadscope, Inc., 1245 Kinnear Road, Columbus, Ohio 43212, and Department of Statistics, The Ohio State University, Columbus, Ohio 43210

Received October 8, 2001

Statistical data mining methods have proven to be powerful tools for investigating correlations between molecular structure and biological activity. Recursive partitioning (RP), in particular, offers several advantages in mining large, diverse data sets resulting from high throughput screening. When used with binary molecular descriptors, the standard implementation of RP splits on single descriptors. We use simulated annealing (SA) to find combinations of molecular descriptors whose simultaneous presence best separates off the most active, chemically similar group of compounds. The search is incorporated into a recursive partitioning design to produce a regression tree for biological activity on the space of structural fingerprints. Each node is characterized by a specific combination of structural features, and the terminal nodes with high average activities correspond, roughly, to different classes of compounds. Using LeadScope structural features as descriptors to mine a database from the National Cancer Institute, the merging of RP and SA consistently identifies structurally homogeneous classes of highly potent anticancer agents.

INTRODUCTION

We present a new method for identifying key structural features of compounds that are related to a measure of biological activity or other property of the compound. This method applies to data sets that are large in terms of both the number of compounds and number of binary descriptors representing the presence or absence of a molecular feature. The method uses multiple features at each splitting node in a recursive partitioning algorithm, coupled with a stochastic search at each node to find a good set of splitting features. We call this method Recursive Partitioning with Simulated Annealing (RP/SA).

Recursive partitioning[1,2] (RP) is a statistical technique that can be useful in attempting to explain a response such as biological activity in terms of a large group of structural features. It can be effective in uncovering structure in data with hierarchical, nonlinear, nonadditive, or categorical variables and has proven useful in classifying pharmaceutical data by discrete or continuous descriptors.[3−7] RP begins with a potentially large heterogeneous set of compounds and attempts to produce smaller and more homogeneous sets. This is accomplished recursively by finding features that successively split a set of compounds into two subsets (+ and − response branches) that are more similar to each other in terms of their biological activity than the original set. In the standard application of RP, each feature can be used to divide the set of compounds into two groups, those with the feature (+) and those without (−). At each step of RP, the feature chosen for the split is the one that creates the most diverse groups according to a criterion such as the difference in mean biological activity of the two groups; alternatively, the criterion for feature selection may be a measure of increased homogeneity within the two subsets that are produced. Formally, the criterion often chosen for variable selection is the *F*-statistic for comparing groups with and without a given feature. Splitting stops when the Bonferonni adjusted P-value for the test fails to achieve significance at some predetermined level. Note that choosing the largest *F*-statistic for comparing two groups is equivalent to choosing the largest *t*-statistic in absolute value, since $F = t^2$.

Young and co-workers[3,4] developed a version of RP called SCAM (Statistical Classification of Activities of Molecules), using a variety of binary 2D and 3D descriptors, each of which represents the presence or absence of a molecular feature. SCAM has been used in a sequential screening algorithm[5] to develop SAR rules from initial screening results and to select additional compounds for subsequent rounds of screening. Chen et al.[7] extended SCAM to identify 3D pharmacophores in the program SCAMPI (Statistical Classification of Activities of Molecules for Pharmacophore Identification). Multiple conformations are examined using a random search technique. This is combined with recursive partitioning to construct more complex pharmacophores from simpler ones; e.g., 3-point pharmacophores from 2-point.

Cho et al.[8] extended recursive partitioning using binary descriptors to identify sets of multiple descriptors for each splitting node. The technique is referred to as BFIRM (Binary Formal Inference-Based Recursive Modeling). An F-test, the ratio of variances between the (+) descriptor and (−) descriptor branches, is the statistical measure for selecting descriptors. Multiple descriptors were selected from among those with the highest individual F-test values. The resulting trees were typically linear; that is, splitting usually gave + branches that were terminal nodes. The authors point out that more useful structural information is obtained when multiple descriptors are used for splitting because compounds in the + branch have a larger common substructure than with single descriptor splits.

* Corresponding author phone: (614)675-3766; e-mail: pblower@leadscope.com.

† Leadscope, Inc.

‡ The Ohio State University.

§ Current address: PhytoCeutica, Inc., 5 Science Park, Suite 13, New Haven, CT 06511.

10.1021/ci0101049 CCC: $22.00
Published on Web 02/05/2002

Miller[9] developed a hybrid approach combining recursive partitioning with K nearest neighbors (KNN) analysis. The combined method takes advantage of the strengths of each component method and each counterbalances a weakness of the other. KNN bases predictions of a test sample on the values of those members of the training sample in its local neighborhood of descriptor space. The strength of RP lies in selecting those descriptors most relevant for classification, while KNN may include irrelevant as well as relevant descriptors.

When a database contains many distinct classes of active compounds, the single best splitting variable in a standard RP algorithm may inadvertently break apart some of these classes. Such is often the case when the variables used in RP do not individually identify single classes. This deficiency may be reduced by first combining variables before making an actual split. Cho et al.[8] adopted this strategy into their BFIRM software. A drawback of their approach, however, is that they use a simple forward selection method for selecting the variables to be used in the splitting.

In the regression context, a greedy forward selection algorithm that chooses the "best" variable to include in the model at each step does not necessarily end up with the best set of variables for the final model. BFIRM, although it does use multiple features for splits, chooses the multiple features for a split based on their ranking individually. Since the best two or three feature set for identifying activity may not involve the best individual features, BFIRM could miss important interactions of this order.

From a conceptual standpoint, the key distinction between our method and BFIRM or standard RP is that both of these methods optimize features individually, whereas we search for optimal combinations of features at each step. The difficulties encountered in choosing the best combination of features for a split are computational. Suppose we are interested in choosing the best combination of three features to use for splitting a node. If there are 1000 features, there are well over 100 million combinations of three features that could be selected from 1000 features, and choosing the best combination of three features could involve comparing over 100 million possible splits. The situation is worse for selecting a best combination of 5 or even 10 multiple features for a split. Our remedy for this is to use a stochastic search algorithm to find a "good" combination of features for a split at any point in the RP algorithm. Regardless of the search algorithm employed, the resulting RP tree will be stochastic. However, despite the fact that the stochastic algorithm produces different trees every time it is used, these trees typically contain many of the same classes of compounds in their upper nodes.

There are a host of stochastic methods that have been proposed to search for points in space at which an objective function is (locally) optimized. The two most popular of these are simulated annealing and the genetic algorithm, both of which are general purpose methods based on natural processes. Simulated annealing[10] emulates the process of alternate heating and slow cooling to achieve low-energy states. A recent application is simulated annealing guided evaluation[11] (SAGE), which is used to identify subsets of compounds with maximal molecular diversity. The genetic algorithm[12] simulates an evolutionary process using randomness of natural selection, crossover, and mutation to approach optimality. It has been used[13] in a related context to select active compounds in a large database by iterating selection and screening cycles.

A novel stochastic procedure is based on the foraging behavior of ants.[14] Most recently, this stochastic technique ANTP has been used[15] to generate stochastic regression trees with the goal of finding a tree with the best cross-validated prediction. In contrast to standard recursive partitioning, which employs a greedy algorithm in choosing splitting variables and cutting values at each iteration, variables are chosen randomly according to weights determined by the previous successes of trees that included these variables. The process not only improves on standard recursive partitioning but also identifies better trees than does a purely random search and in a much shorter time.

The stochastic modification of recursive partitioning that we propose is different from those used in SCAMPI and ANTP in several important ways. Most importantly, if the splits are based on K features (with K in the range of 2−5) from a data set with M features, the stochastic search is over all (M choose K) combinations of 2D structural features from a feature set with M typically in the thousands. The chosen combination results in a single split in the RP-tree, which roughly identifies a meaningful group of compounds. The order and manner in which such groups are split from the main trunk of the tree is not important. In particular, due to possible aliasing, several different subsets of 2D structural features may identify essentially the same group of compounds. The main point is that four or five 2D structural features are typically sufficient to identify a biologically active group of compounds, so the search space should follow this phenomenon.

Although the BFIRM procedure uses multiple binary descriptors in each split of its tree, these are selected greedily, which may miss the subtle types of interactions that need to be identified. In contrast, RP/SA has the ability to discover many unanticipated combinations, and its method of establishing an overall noise level obviates the need for cumbersome pruning rules. In the well studied NCI data set, RP/SA not only identified the well-known classes of compounds active against the NCI−H23 cell line but also discovered several less well-known classes as well.

METHODS

The RP/SA algorithm is composed of two separate algorithms; RP to classify the data and SA to determine which molecular descriptors best partition each node of the tree. In particular, the SA algorithm searches for a combination of *K* features to maximize a particular splitting criterion. For example, the splitting criterion might be the difference in the means of the two groups, a *t*-statistic comparing the two groups, or a maximization of some other measure of within group homogeneity.

Splitting Criterion. The default splitting criterion is the P-value of the two-tailed *t*-statistic for the difference in mean response between the two subsets of compounds that would be produced by the proposed split. This is the standard criterion and the only option available in some RP packages. Its justification is based on the approximately normal distribution of sample means as well as computationally efficient updating.[4] Typically, a minimum set size must also be specified in order to disregard aberrant observations.

An apparently similar criterion is to choose the variable producing the largest absolute difference, D^*, in means between the two subsets. The more standard *t*-criteria chooses the largest value t^* of

$$t = \left(\sqrt{n_+\left(\frac{n_-}{n_+ + n_-}\right)}\right)\left(\frac{D}{s_p}\right)$$

where D is the absolute difference between the two means, s_p is the pooled standard deviation, and n_+ and n_- are the sample sizes for the positive and negative branches, respectively. When $n_- \gg n_+$, which typically occurs in a large node with a small proportion of active compounds, then

$$t \approx \sqrt{n_+}\frac{D}{s_p} \approx \sqrt{n_+}\frac{D}{\sigma}$$

where σ is the standard deviation of the node being split. Thus, the *t* criterion does not necessarily split on the largest mean difference but rather the largest value of $\sqrt{n_+}D$. This tends to favor splits with larger values of n_+ and possibly smaller values of D.

The purpose of the *t* criterion is to avoid splitting a node into daughters that differ just by chance. The largest *t* corresponds to the smallest P-value and the most "statistically significant" difference. However, in many of our examples we find the P-values associated with both the *t* criterion and the largest mean difference criterion to be less than 10^{-10}, so arguments for finding the smallest P-value as opposed to the largest mean difference on the basis of statistical significance would seem less compelling. In fact, in the examples in the results section it will be seen that the mean criterion tended to produce simpler trees with higher activities in the terminal nodes.

A key to the successful use of the mean criterion is to work with an appropriate minimum node size. This will depend on the application as well as the choice of features and compounds. If small minimum node sizes are allowed, then a minimum P-value criterion, such as 10^{-5}, should be included as well to avoid spurious splits.

Simulated Annealing Portion of RP/SA. Initialization: 1. The algorithm is initialized with a temperature $T = T_0$ and a number K of features that will be used to determine the split according to the chosen splitting criterion. A parameter n representing the minimal number of allowable compounds in any node must also be specified, although this parameter can decrease depending on the depth level of the node in the tree. The chemical space of features is then filtered to include only those features that appear in at least n compounds, and this is the feature pool.

2. A set of K features is chosen randomly from the feature pool. A query is made to determine which compounds in the dataset contain the simultaneous presence of these K features. If the number of compounds containing the K features is smaller than the frequency cutoff n, another random selection is made until the number of compounds is larger than n. Having determined the set A of compounds that contain this combination of K features, we calculate the value μ_A of the splitting criterion.

Iteration: 3. First a random number $J \in (0, J_{Max}]$ is chosen, and J features are dropped at random from the current set of K features. Next a random sample of J features is selected from the feature pool. These replace the J features that were dropped.

The maximum size J_{Max} of the J features allowed is a function of the temperature T, and it gets smaller as T decreases. In this work the following schedule was used for J_{Max}:

$$J_{Max} = \begin{cases} K - 1 & \text{if } T > 10^{-1} \\ \lfloor K/2 \rfloor & \text{if } 10^{-2} \le T \le 10^{-1} \\ 1 & \text{if } T < 10^{-2} \end{cases}$$

The J features are chosen repeatedly until the minimum node size n criterion is satisfied. For the resulting set B of compounds, we calculate the value μ_B of the splitting criterion.

4. If $\mu_B \ge \mu_A$, the K features, with the new subset of J features, are accepted and used as the current set of K features; the temperature is decreased, $T_{new} = \alpha T$ where $\alpha \in (0,1)$; and the procedure returns to step (3).

5. If $\mu_B < \mu_A$, a random $R \in [0,1)$ is selected and a check is made against the Metropolis condition

$$R < \exp\left(-\frac{(\mu_A - \mu_B)}{(b_1 - b_0)T}\right)$$

where b_0 is the mean activity of the full set and b_1 is the threshold activity for determining active compounds. If this condition is met, the K features with the new subset of J features are accepted; the temperature is decreased to $T_{new} = \alpha T$; and the procedure returns to step (3).

6. If the two criteria in (4) and (5) above are not met, steps (3)−(5) are repeated. If the procedure does not satisfy the criteria of steps (4) and (5) for *100* times, the temperature is lowered.

7. The process terminates when the current temperature falls below a minimum temperature cutoff, T_{min}.

When using SA on the NCI dataset, the input parameters to the SA algorithm were chosen as the following: $T_0 = 10$, $T_{min} = 10^{-3}$, $\alpha = 0.9$. The term $b_1 - b_0$ adjusts for the scale of the activity data, while the choice of T_0 and α will depend on the smoothness of the response surface. The minimum frequency cutoff n was varied between 10 and 100 during different runs. Although the number K of features to consider in combination is arbitrary in principle, typically K is chosen to lie in the range [2,5].

Recursive Partitioning Portion of RP/SA. This procedure is shown as a flowchart in Figure 1. The SA algorithm is used to split the data set at each level in an RP tree. Figure 2 shows a typical result of the RP/SA algorithm for a combination of four features, using the difference in mean activity as the splitting criterion and a minimum frequency cutoff of $n = 50$. A dataset of $N = 28\,297$ compounds with a mean pGI_{50} activity of 4.47 represents the root node. SA is used to identify a combination of four features that are associated with activity. At the first split, SA identifies a set of four features that must be present: *ether, aryl-; hacceptor-path3-hacceptor; sulfide, alkenyl-* and *benzene, 1-(alkenyl, cyc)-*. This split separates the compounds into two groups: one group of 51 compounds with an average pGI_{50} value of 7.74 that contains this particular combination of features and the other group is the complement set. The (+) branch is a

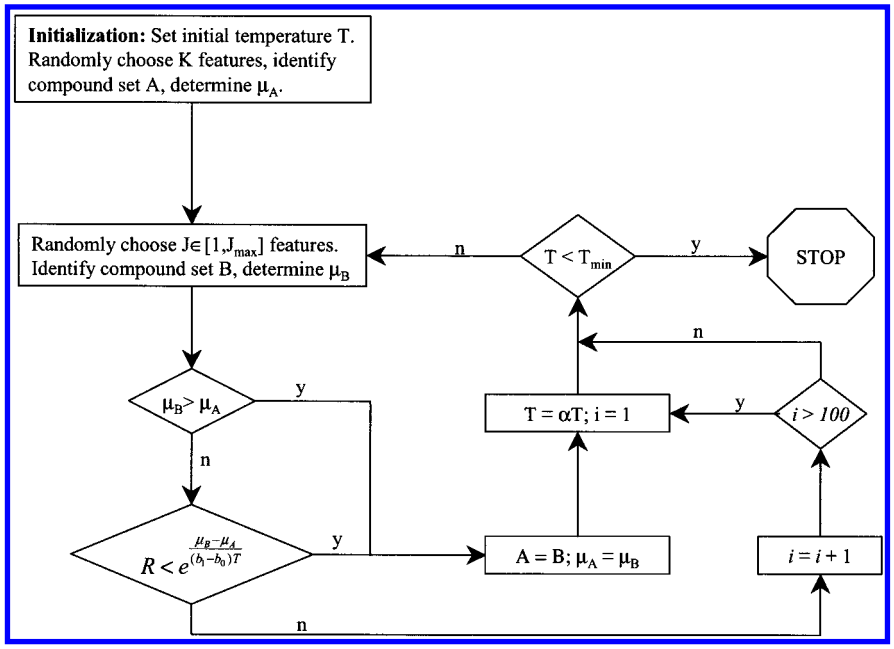

Figure 1. Simulated annealing algorithm. *A* and *B* denote subsets of *K* molecular descriptors and also the corresponding compound sets; *i* is a counter which records the number of iterations since *T* was lowered; b_0 is the mean activity of the full set; and b_1 is the threshold activity for determining active compounds. J_{Max} decreases with *T* according to the schedule described in the text.

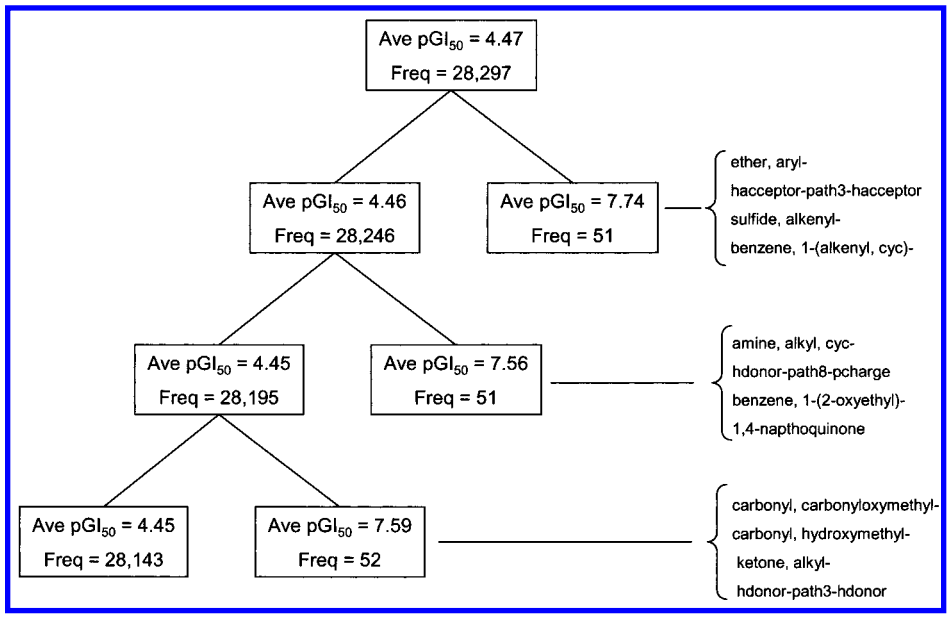

Figure 2. A sample RP/SA tree with three levels; parameters: four features in combination, minimum node size = 50. Feature sets are listed for the (+) branches; mean pGI_{50} value and node size are shown for all nodes.

terminal node because of the minimum frequency requirement of 50. Reapplying SA on the group of 28 246 compounds identifies a second four-feature combination of *amine, alkyl, cyc-; hdonor-path8-pcharge; benzene, 1-(2-oxyethyl)-* and *1,4-naphthoquinone*. The procedure is performed recursively until the depth of the tree reaches some predefined level set by the user.

Stopping Criteria. In its current version, RP/SA terminates by the user specifying a maximum depth to which to grow the tree. This may be determined either by practical limitations of time and complexity or by using the following statistical guideline for stopping the partitioning process. We first scramble the compounds by randomly permuting their activities to eliminate any natural association between structural features and activity that may have been present in the data. We then run RP/SA for the scrambled data to the maximum number of levels anticipated. The mean activity of the more active node is computed at each level. This is illustrated by the broken line in Figure 3, which represents a general noise level that should remain fairly constant over the levels of the tree. For the actual data, the RP/SA tree is grown until the mean activity of the more active node approaches the level of noise.

Run times depended on the partitioning technique and, for RP/SA, the number of features K used in combination. On a Pentium 3 with a 1 GHz processor and 512MB of memory, recursive partitioning of the NCI database took 0.66 s/node. RP/SA with $K = 2$ took 9.4 s/node, while RP/SA with $k = 4$ took 18.0 s/node.

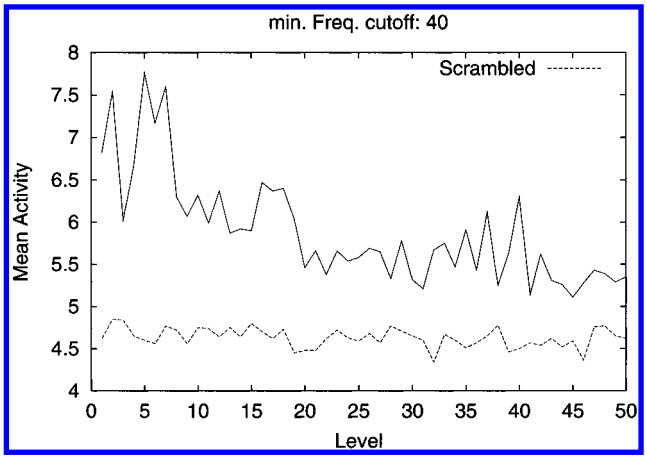

Figure 3. The mean activity (pGI_{50}) plotted versus the level of a RP/SA tree for the NCI−H23 data. The lower curve is a RP/SA tree with the activities randomly assigned to all the compounds in the NCI dataset (scrambled).

RESULTS

NCI Database. The database used in this study was obtained from the National Cancer Institute's (NCI) Developmental Therapeutics Program.[16] Since 1990, the DTP Human Tumor Cell Line Screen has been testing compounds for growth inhibition against a panel of 60 human tumor cell lines[17] in microtiter plate format. For each compound and cell line, growth inhibition after 48 h of drug treatment is assessed from changes in total cellular protein using a sulforhodamine B assay.[17] The data provide three concentration parameters for each compound-cell line pair: the GI_{50} value is the concentration that causes 50% growth inhibition; the TGI value is the concentration needed for "total growth inhibition"; and the LC_{50} value is the lethal concentration at 50%. The example below uses GI_{50} data for the NCI−H23 cell line from the nonsmall cell lung cancer panel, which are provided by the DTP as −log(1/C), denoted pGI_{50}. There were 28 297 compounds with measured values of pGI_{50} for the NCI−H23 cell line. The average of these values is $b_0 = 4.47$. We chose $b_1 = 5.47$ as the activity threshold of interest, giving a convenient SA scaling factor ($b_1 - b_0$) of 1.

Molecular Descriptors. The molecular descriptors used in this study are those defined in the LeadScope Structural Feature Hierarchy.[18] This is based on structural features and combinations of features commonly used for experimental design in drug discovery programs—the building blocks of medicinal chemistry. When LeadScope loads a set of compounds to create a project, the software performs a systematic substructure analysis using predefined structural features stored in a feature library. The structural features chosen for analysis are motivated by those typically found in small molecule drug candidates. At the present time, the feature library contains over 27 000 structural features.

The major structural classes are as follows: *amino acids; bases, nucleosides; benzenes; naphthalenes; carbocycles; carbohydrates; elements; functional groups; heterocycles; natural products; peptidomimetics; pharmacophores; protective groups* and *spacer groups*. The features represent a wide range of structural specificity from very specific substructures such as *benzene*, *1-hydroxymethyl*, *3-methoxy-* to generic features such as the *pharmacophores* which are pairs of generalized physiochemical atom types joined by a path of

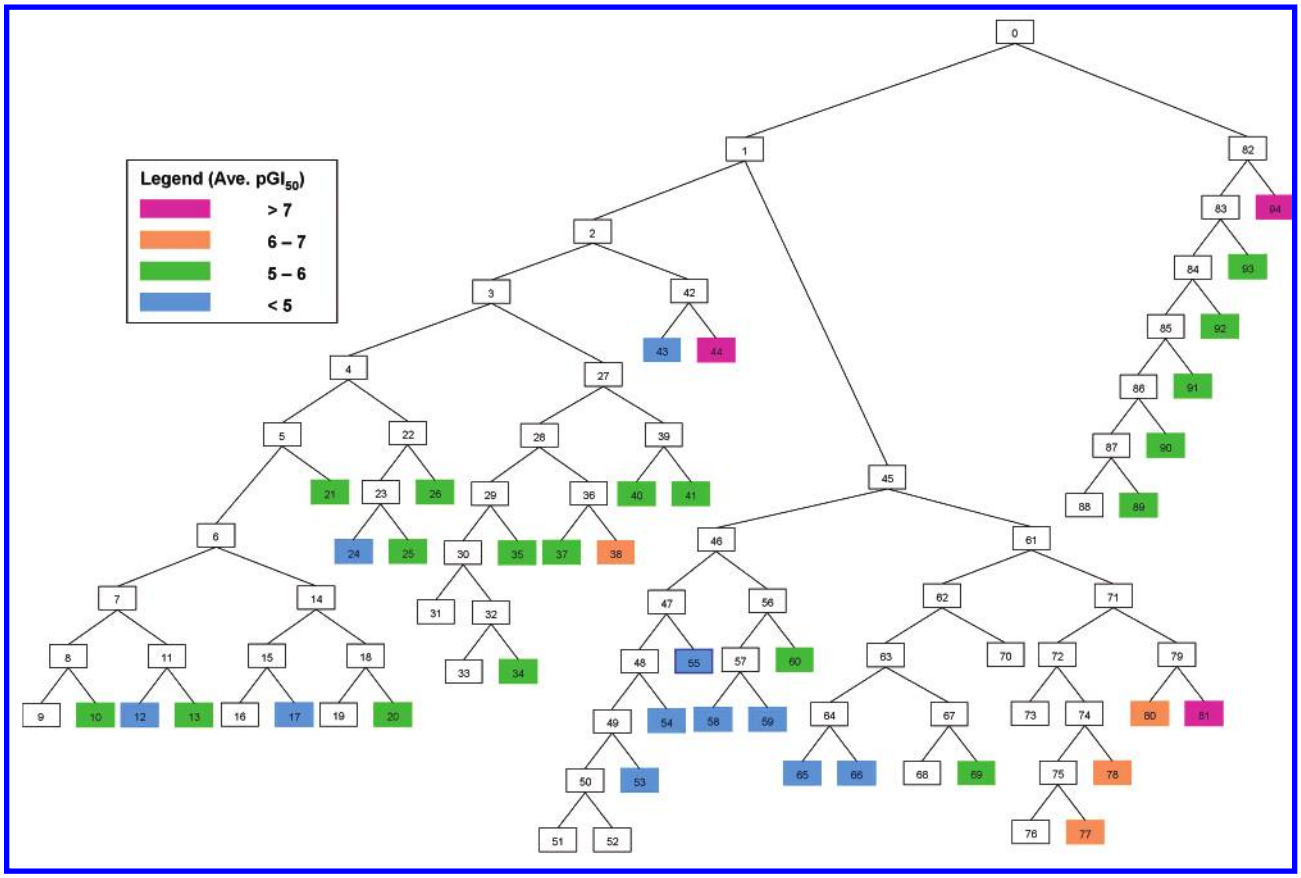

Figure 4. Recursive partitioning tree for the NCI−H23 data; parameters: splitting criteria = *t*-test, maximum P-value = 0.01, minimum node size = 50. Node details are given in Table 1. Terminal nodes are highlighted according to the legend.

Table 1. Node Details for RP Tree in Figure 4[a]

node	splitting feature	mean activity	node size	node	splitting feature	mean activity	node size
0	root	4.47	28297	48	(pyridine(H), 2-aryl)	4.23	729
1	(1,4-benzoquinone)	4.44	27521	49	(benzene, 1-(p-alkyl)-)	4.18	654
2	(oxycarbonyl, O-alkyl, cyc)	4.40	25408	50	(cyclohexene)	4.15	574
3	(tropone)	4.39	25274	51	(oxolane)	4.06	396
4	(pyridine, 3-fused ring)	4.36	23583	52	oxolane	4.36	178
5	(organometal)	4.35	23304	53	cyclohexene	4.43	80
6	(pyridine, 2-hydrazonomethyl)	4.34	23177	54	benzene, 1-(p-alkyl)-	4.66	75
7	(benzene, 1-hydroxy-)	3.97	23225	55	pyridine(H), 2-aryl	4.74	53
8	(pyran(H), 3-methyl-)	4.31	21225	56	carboxylate, alkenyl	4.88	237
9	(pyridinium, 1-fused ring)	4.31	21142	57	(pyran(H)))	4.70	159
10	pyridinium, 1-fused ring	5.34	83	58	(oxolane, 3-fused ring)	4.46	83
11	pyran(H), 3-methyl-	5.22	125	59	oxolane, 3-fused ring	4.96	76
12	(hacceptor-path3-hacceptor)	4.51	61	60	pyran(H)	5.25	78
13	hacceptor-path3-hacceptor	5.90	64	61	hacceptor-path8-hdonor	5.36	1094
14	benzene, 1-hydroxy-	4.58	1827	62	(carbonyl, alkyl, cyc)	4.74	509
15	(aromatic-path8-hdonor)	4.41	980	63	(alkene, trisubst)	4.57	392
16	(ketone, alkenyl)	4.38	885	64	(alcohol)	4.41	216
17	ketone, alkenyl	4.72	95	65	(oxolane, 2(alkyl, acyc)-)	4.17	98
18	aromatic-path8-hdonor	4.77	847	66	oxolane, 2(alkyl, acyc)-	4.61	118
19	(sec-amine(NH), phenyl)	4.72	792	67	alcohol	4.77	176
20	sec-amine(NH), phenyl	5.46	55	68	(aromatic-path5-hacceptor)	4.63	112
21	pyridine, 2-hydrazonomethyl	5.89	127	69	aromatic-path5-hacceptor	5.02	64
22	organometal	5.39	279	70	alkene, trisubst	5.32	117
23	(tin, p-alkyl)	5.12	187	71	carbonyl, alkyl, cyc	5.90	585
24	(benzene, monosubst-)	4.85	94	72	(amine, alkyl, acyc-)	5.51	410
25	benzene, monosubst-	5.39	93	73	(hacceptor-path3-hacceptor)	4.77	106
26	tin, p-alkyl	5.94	92	74	hacceptor-path3-hacceptor	5.76	304
27	pyridine, 3-fused ring	4.85	1691	75	(1,3-diene)	5.51	239
28	(sec-amine(NH), p-alkyl)	4.77	1497	76	(tetralin)	5.21	178
29	(hdonor-path7-pcharge)	4.70	1389	77	tetralin	6.38	61
30	(amine, dimethyl, alkyl)	4.67	1334	78	1,3-diene	6.67	65
31	(pyridine, 2-aryl-)	4.60	1065	79	amine, alkyl, acyc-	6.83	175
32	pyridine, 2-aryl-	4.94	269	80	(carbonyl, aryl-)	6.38	99
33	(1,8-naphthyridine)	4.78	211	81	carbonyl, aryl-	7.42	76
34	1,8-naphthyridine	5.55	58	82	1,4-benzoquinone	5.36	776
35	amine, dimethyl, alkyl	5.47	55	83	(benzene, 1-alkoxymethyl)	5.17	700
36	hdonor-path7-pcharge	5.67	108	84	(tert-amine)	5.07	601
37	(benezene, 1,2-fused,4-acyc)	5.17	54	85	(ether, alkenyl-)	4.99	535
38	benezene, 1,2-fused,4-acyc	6.17	54	86	(alcohol, alkyl-)	4.91	475
39	sec-amine(NH), p-alkyl	5.50	194	87	(aromatic-path3-pcharge)	4.86	402
40	(acridine)	5.01	94	88	(carbonyl, alkyl, acyc-)	4.80	335
41	acridine	5.96	100	89	carbonyl, alkyl, acyc-	5.15	67
42	tropone	6.30	134	90	aromatic-path3-pcharge	5.20	73
43	(benzene, 1-alkenyl, 2-alkoxy)	4.91	62	91	alcohol, alkyl-	5.60	60
44	benzene, 1-alkenyl, 2-alkoxy	7.50	72	92	ether, alkenyl-	5.71	66
45	oxycarbonyl, O-alkyl, cyc	4.92	2113	93	tert-amine	5.77	99
46	(hacceptor-path8-hdonor)	4.45	1019	94	benzene, 1-alkoxymethyl	7.08	76
47	(carboxylate, alkenyl)	4.32	782				

[a] Features listed in parentheses represent nodes for which the feature is absent.

atoms/bonds of indeterminant type. Pharmacophores are very similar to the binding property pairs of Kearsley et al.[19]

Standard Recursive Partitioning Tree. Figure 4 shows the tree that results from running the standard recursive partitioning algorithm down to a depth of 10 with a minimum node size of 50 compounds and Table 1 gives details about the nodes. Internal nodes appear in white, and terminal nodes are colored coded by the average (pGI_{50}) activity. For improved appearance, once the tree reaches depth 7, it omits the daughters of a few internal nodes. All of the 35 terminal nodes displayed had "significantly" ($P < 0.0001$) higher means than their sister nodes, except when both sisters were terminal. The average activity level varies among these 36 terminal nodes as follows: 3 nodes have very high (>7) average activity, 4 have high (6−7), 18 have moderate (5−6), and 11 have low (<5) average activity. Thus, for this example, the standard RP tree is quite complex and only modestly successful at identifying active groups of compounds.

To quantify the effectiveness of a tree in identifying active groups of compounds, we use the average activity of the 10 positive nodes first encountered in growing the tree. More precisely, we first determine the depth d at which the tenth positive terminal node is encountered. Let a_1 be the average activity of all positive terminal nodes occurring above depth d; let a_2 be the average activity of all positive terminal nodes occurring at depth d; and let c be the number (<10) of all positive terminal nodes occurring above depth d. Then our measure of "quickly found" activity is

$$Q = (c/10)a_1 + (1 - (c/10))a_2$$

One reason for this measure is to have a fairer comparison between trees that tend to be bushy such as the RP tree versus the more linear trees produced by the other methods presented. Since the activity of the terminal nodes tends to decrease somewhat with the depth of the tree, averaging all the terminal nodes of the RP tree would give the other methods an unfair advantage if all trees were grown to the same depth. Since some trees can go down to a depth of 80 or more, this also has the advantage of not having to grow the tree that far for comparison purposes.

Figure 5 contains plots of Q versus the minimum number of compounds in a terminal node, for four different tree-growing algorithms. The absolute upper bound of Q is found by first sorting all compounds by their activity. For a minimum node size of 25, since there are 10 levels we average the activities of the $10 \times 25 = 250$ highest

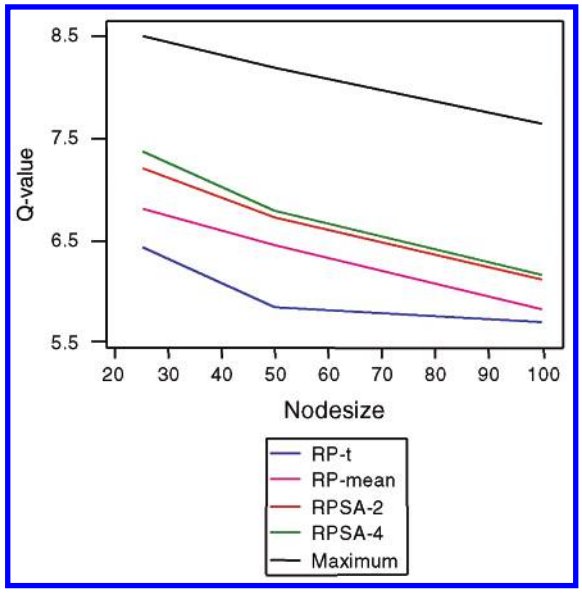

Figure 5. A comparison of *Q* values for the recursive partitioning and RP/SA trees. Q measures the average activity for the top 10 terminal nodes.

compounds to obtain the absolute upper bound for Q. It is not possible to achieve or come close to this upper bound as this would require the top 250 compounds fall into 10 disjoint groups of 25 structurally similar compounds which can be identified by the features. The usefulness of the upper bound pertains to its rate of decrease, which is very close to the rate found for each of the methods. For standard recursive partitioning, the values of Q corresponding to minimum node sizes of 25, 50, and 100 are 6.42, 5.84, and 5.68, respectively. This quantifies the general impression given by Figure 4 that active nodes appear only sporadically around the standard RP tree.

RP Tree Using Mean Difference Criterion. A simple adjustment to the standard RP tree is to replace the *t*-statistic criterion for splitting a node with the simpler criterion of the difference in means. For the NCI database coupled with our feature set, this leads to a remarkable improvement in detecting the more active groups of compounds. In Figure 5, the *Q* values for this procedure range about 5% higher than those for standard RP. Moreover, as seen in Figure 6, the tree resulting from RP using the mean criterion and a minimum set size of 50 is quite a bit simpler than the tree from the standard RP algorithm as shown in Figure 4. Table 2 gives details about the nodes of this tree. The same kind of simplification occurs for other node sizes, although for our feature set the tree remains somewhat bushy for a minimum node size of 25.

RP/SA Trees. Unlike the RP trees that produce a single tree whether using the *t* or the mean criterion, RP/SA has a stochastic component and thus produces a different tree with each run. Because of this, we have generated 10 trees for any given set of conditions. The randomness is beneficial in exploring different solutions and determining which parts of the solutions are stable. In particular, when growing a tree to 10 levels, the average activity of the terminal nodes (Q value) was extremely stable over the 10 trees produced. Figure 7 illustrates the best and worst *Q* value solutions over the 10 trees that were grown using two features and a

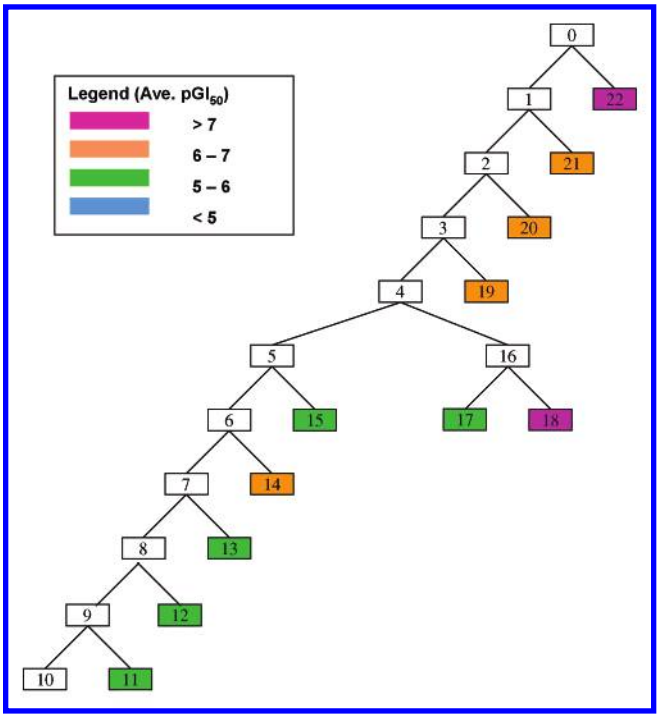

Figure 6. Recursive partitioning tree for the NCI−H23 data; parameters: splitting criteria = maximum mean pGI_{50}, minimum node size = 50. Node details are given in Table 2. Terminal nodes are highlighted according to the legend.

Table 2. Node Details for RP Tree in Figure 6[a]

node	splitting feature	mean activity	node size
0	root	4.47	28297
1	(oxetane,2-(alkyl,cyc))	4.46	28244
2	(benzene, 1-carbonyl, 4-(2-oxyethyl))	4.45	28164
3	(quinoline, 2-(alkenyl, cyc))	4.45	28103
4	(oxepin, 4-oxymethyl)	4.45	28050
5	(tropone)	4.44	27906
6	(oxolane, 3-aminomethyl-)	4.43	27815
7	(quinoline, 5-nitro)	4.43	27763
8	(1,4-benzoquinone, 2-alkoxy-)	4.43	27699
9	(tin, p-alkyl)	4.42	27597
10	(isoquinoline, 8-methyl)	4.42	27527
11	isoquinoline, 8-methyl	5.84	70
12	tin, p-alkyl	5.9	102
13	1,4-benzoquinone, 2-alkoxy-	5.9	64
14	quinoline, 5-nitro	6.03	52
15	oxolane, 3-aminomethyl-	5.99	91
16	tropone	6.23	144
17	(benzene, 1,3-dimethoxy)	5.26	84
18	benzene, 1,3-dimethoxy	7.6	60
19	oxepin, 4-oxymethyl	6.38	53
20	quinoline, 2-(alkenyl, cyc)	6.79	61
21	benzene, 1-carbonyl, 4-(2-oxyethyl)	6.86	80
22	oxetane,2-(alkyl,cyc)	7.1	53

[a] Features listed in parentheses represent nodes for which the feature is absent.

minimum node size of 50. The figure indicates the lowest and highest *Q* values are 6.67 and 6.78, respectively, while the *Q* value for the nonrandom tree in Figure 6 using a single feature had an associated *Q* value of 6.44. The improvement going from an exhaustive search of 1 feature to a random search of two features is about 5%, with essentially no chance of one of the random two feature solutions having a *Q* value lower than the one feature solution. In addition, a visual comparison of Figure 7 with either Figures 6 or 4 shows that the individual nodes found by RP/SA are more consis-

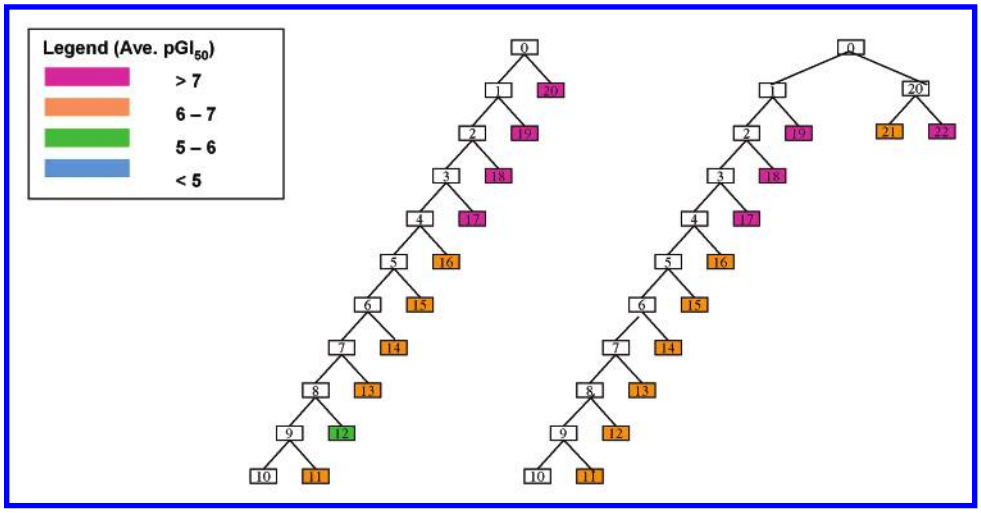

Figure 7. Sample RPSA trees for the NCI−H23 data. RPSA parameters: $K = 2$ features and minimum node size = 50. Node details are given in Table 3. Terminal nodes are highlighted according to the legend. Tree 1 (left) − average pGI_{50} in terminal nodes is 6.67; tree 2 (right) − average pGI_{50} in terminal nodes is 6.78.

tently active than those found by either of the other deterministic methods.

As seen in Figure 5, the improvement in *Q* values when using RP/SA persist for the other two minimum node sizes considered. For our feature set and the NCI data, very little was gained by using four features as opposed to two. This may not occur for feature sets that have more prevalent features, since in this case more features may be necessary to identify active subsets of moderate size.

The major source of consistency across stochastic trees arises from their tendency to capture many of the same compounds. Despite the randomness, the 10 RP/SA trees with two features and minimum node size of 50 have a great degree of commonality in this regard. To measure this, consider the 1315 different compounds that appeared in at least one of the 10 trees. The commonality of each of these compounds is defined as the number of trees (1−10) in which it appeared. Table 4 shows the number of compounds with each level of commonality as well as the average activity of these compounds. Approximately 32% (426/1315) of the compounds found appear in only one tree, and the average activity of these compounds is only 5.7. At the other extreme, approximately 20% of the compounds found appeared in all 10 trees and these had a high average activity of 7.1. There is a consistent trend in which the highly active compounds that are found at all are found repeatedly. The regression line illustrating this highly significant trend is shown in Figure 8.

Consider again using RP/SA with two features and a minimum node size of 50. Although the average *Q* value is fairly stable over the 10 stochastic trees, the average activity for the compounds in the first level of the trees is quite variable and ranges from 5.91 to 7.61. An exhaustive search would have an average activity of at least 7.61 at the first level. Would the procedure do better overall if a more thorough search were conducted at each node? To investigate this, we grew two "super" trees, where at each level we ran the simulated annealing 10 times and chose the best of the 10 solutions. The *Q* values for these trees were 6.80 and 6.74 and were both comparable to the values found for the regular RP/SA trees. Thus, even when using multiple features a greedy stepwise procedure does not necessarily lead to a better overall solution. Nevertheless, optimizing with respect to two features at each step produces more active groups than the use of a single feature at each step.

DISCUSSION

Although RP/SA is designed to identify groups of active compounds having a few key features in common, an additional benefit is that RP/SA tends to find active groupings having many additional common features. Such groupings may be recognized as structural classes or subclasses. When active classes or subclasses are present in the database, they have a higher probability of being found during the SA cycle of RP/SA simply because many different combinations of features can be used to include the active class or subclass.

Researchers at the National Cancer Institute and elsewhere have identified several classes of highly active anticancer agents contained in the NCI database.[20,21] These include actinomycins, anthracylcines, campothecins, podophyllotoxins, taxanes, colchicines, and macrolides (see Figure 9). The compounds in these highly active classes are consistently found in the first 10 tree levels generated by RP/SA for the NCI−H23 cell line data. In addition, the other well-known classes of anticancer agents such as ellipticines, 9-aminoacridines, and thiosemicarbazones of 2-carbonyl-pyridines (see Figure 9) are also found repeatedly. But the latter classes have somewhat lower activity and are not found as consistently.

Table 5 summarizes the distribution of common classes of anticancer agents in 100 nodes from the first 10 levels of ten RP/SA trees generated with combinations of two features and minimum node size of 50. Note that six of the common classes of anticancer agents shown in Figure 9 appeared in 10 of 10 RP/SA trees, two classes appeared in six of 10 trees, and two appeared in four trees. The column labeled *Node Size* is the average size of the tree node containing the class listed, *Class Size* is average number of class members in the node, and *Mean Activity* is the average pGI_{50} value. All statistics are averages over the number of tree nodes in which the class appeared. Thus, the 10 nodes containing *anthracyclines* on average had a pGI_{50} value of 7.43 and contained 54.5 compounds of which 50.5 belonged to that class. Six

Table 3. Node Details for the Two RP/SA Trees in Figure 7: (a) Tree 1 and (b) Tree 2[a]

node	splitting feature	mean activity	node size
	(a) Tree 1		
0	root	4.47	28297
1	(amine, allyl-; benzene, 1-alkyl-, 4-methoxy-)	4.46	28195
2	(oxetane, 2-(alkyl, cyc)- -; hacceptor-path5-hacceptor)	4.45	28144
3	(benzene, 1-carbonyl, 4-(2-oxyethyl);hdonor-path8-pcharge)	4.45	28092
4	(quinoline, 3-(alkyl, cyc)-; pyridine(H), 4-oxymethyl-)	4.44	28039
5	(oxepin, 3-oxymethyl-; pyran(H), 2-fused ring-)	4.44	27988
6	(benzene, 1-methoxy-,3-(3-oxopropyl)-; oxolane, 3-aminomethyl-)	4.43	27916
7	(hacceptor-path6-pcharge; indole, 5-fused ring-)	4.43	27864
8	(pyridine, 4-alkylamino-; benzene, 1,2-fused)	4.43	27797
9	(pyran(H), 2-methyl-; alcohol, t-alkyl-)	4.42	27732
10	(pyridine, 2-hydrazonomethyl-; semicarbazone, thio)	4.42	27624
11	pyridine, 2-hydrazonomethyl-; semicarbazone, thio	6.1	108
12	pyran(H), 2-methyl-; alcohol, t-alkyl-	5.9	65
13	pyridine, 4-alkylamino-; benzene, 1,2-fused	6.21	67
14	hacceptor-path6-pcharge; indole, 5-fused ring-	6.1	52
15	benzene, 1-methoxy-,3-(3-oxopropyl)-; oxolane, 3-aminomethyl-	6.33	72
16	oxepin, 3-oxymethyl-; pyran(H), 2-fused ring-	6.65	51
17	quinoline, 3-(alkyl, cyc)-; pyridine(H), 4-oxymethyl-	7.11	53
18	benzene, 1-carbonyl, 4-(2-oxyethyl);hdonor-path8-pcharge	7.49	52
19	oxetane, 2-(alkyl, cyc)- -; hacceptor-path5-hacceptor	7.22	51
20	amine, allyl-; benzene, 1-alkyl-, 4-methoxy-	7.61	102
	(b) Tree 2		
0	root	4.47	28297
1	(1,4-naphthoquinone,5-hydroxy; alcohol,alkyl-)	4.46	28184
2	(oxetane, 3-oxy-; alkene)	4.45	28133
3	(benzene, 1-alkenyl-, 3-alkoxy-; sulfide, alkyl, cyc-)	4.45	28082
4	(quinoline, 3-fused ring-; pyridine(H), 4-oxymethyl-)	4.44	28029
5	(amino(NR)carbonyl, *N*-methyl-; carboxamide(NHR), alkyl,-cyc-)	4.44	27976
6	(oxepin, 3-oxymethyl-; pyran(H), 2-fused ring-)	4.43	27925
7	(oxolane, 3-aminomethyl-; benzene, 1-alkoxy-, 3-methoxy-)	4.43	27853
8	(pyridine, 2-hydrazonomethyl-; semicarbazone, thio)	4.42	27745
9	(hacceptor-path6-pcharge; indole, 6-fused ring-)	4.42	27692
10	(benzene, 1,2-fused; pyridine, 4-alkylamino-)	4.41	27625
11	benzene, 1,2-fused; pyridine, 4-alkylamino-	6.21	67
12	hacceptor-path6-pcharge; indole, 6-fused ring-	6	53
13	pyridine, 2-hydrazonomethyl-; semicarbazone, thio	6.1	108
14	oxolane, 3-aminomethyl-; benzene, 1-alkoxy-, 3-methoxy-	6.33	72
15	oxepin, 3-oxymethyl-; pyran(H), 2-fused ring-	6.65	51
16	amino(NR)carbonyl, *N*-methyl-; carboxamide(NHR), alkyl,-cyc-	6.96	53
17	quinoline, 3-fused ring-; pyridine(H), 4-oxymethyl-	7.11	53
18	benzene, 1-alkenyl-, 3-alkoxy-; sulfide, alkyl, cyc-	7.74	51
19	oxetane, 3-oxy-; alkene	7.22	51
20	1,4-naphthoquinone,5-hydroxy; alcohol,alkyl-	6.68	113
21	(hdonor-path8-pcharge; benzene, 2-oxyethyl)	6.01	62
22	hdonor-path8-pcharge; benzene, 2-oxyethyl	7.48	51

[a] Features listed in parentheses represent nodes for which the feature is absent.

Table 4. Distribution of Commonality and Associated Mean Activity for the RP/SA Trees with Two Features and Minimum Node Size of 50

commonality level	1	2	3	4	5	6	7	8	9	10
no. of compounds	426	125	137	131	37	47	61	15	69	267
mean activity	5.7	5.9	6.0	6.2	6.3	6.5	6.1	7.7	6.8	7.1

of 10 trees identified *actinomycins* with an average pGI_{50} value of 6.96. The average node contained 58 compounds; 20 were *actinomycins*. Only 20 compounds were *actinomycins* so each of the six trees identified all members of the class. Although the tree nodes containing the classes listed in Table 6 also contained compounds that were not members of the class, the classes were cleanly separated from each other.

For comparison, Table 6 summarizes the distribution of common anticancer classes in the nodes of the RP tree of Figure 4 generated using the *t*-statistic as splitting criteria and minimum node size of 50. The column labeled *Node* identifies the number of the node in Figure 4 containing the class. The columns labeled *Node Size*, *Class Size*, and *Mean Activity* are analogous to those in Table 5, except that reported numbers are observed from a single tree rather than averages over several trees. In some cases, RP was more effective in identifying the common classes of anticancer agents. For example, node 94 contained 63 *anthracyclines* with average pGI_{50} value of 7.08, and node 44 contained 62 *colchicines* with average pGI_{50} value of 7.5. However, some classes were not cleanly separated from each other. For

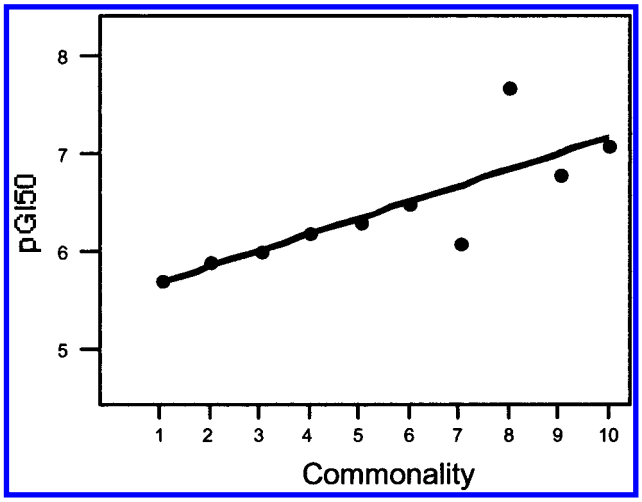

Figure 8. Mean activity vs commonality for the RP/SA trees with two features and minimum node size of 50.

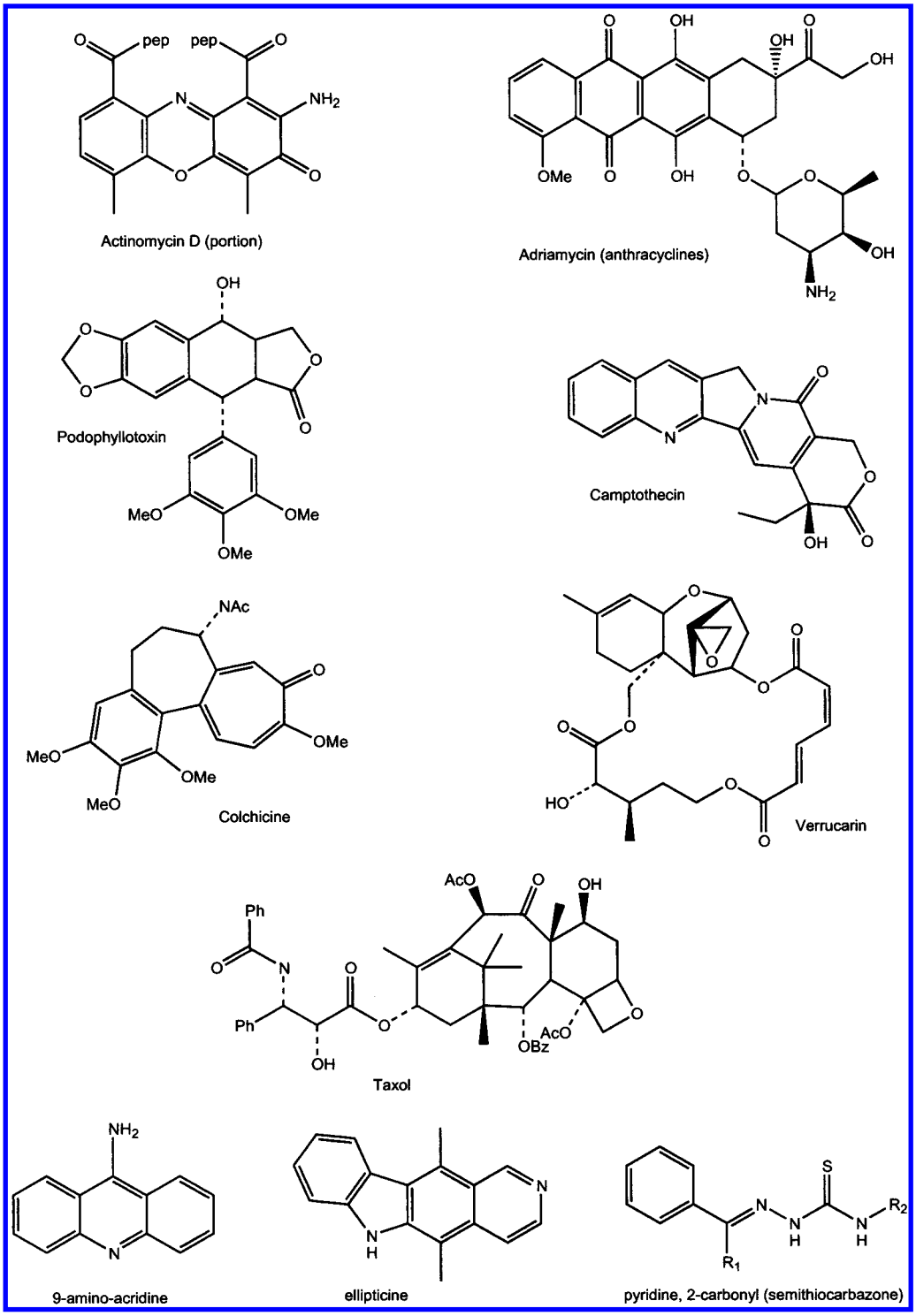

Figure 9. Representatives of well-known classes of anticancer agents.

example, node 78 contains the largest group (28) of *camptothecins* and the only group of *macrolides*; the other seven *macrolides* were scattered as singletons among other nodes. Similarly, node 81 contains 43 *taxanes* and 17 *actinomycins*.

RP/SA can sometimes identify especially active subclasses within established cancer classes. The colchicines provide a good example of this. The NCI dataset contains 78 colchicines with an average activity of 7.0 against the NCI−H23 cell line. However, in six of the 10 trees listed in Table 5, RP/SA identifies a subclass of 38 analogues with an average pGI_{50} value of 7.88. The structural characteristics of this subclass are described in the literature[22] and are summarized in Figure 10a. All members have a thiomethyl substituent on the tropone ring replacing the methoxy group of the parent compound, and all have 1,2-dimethyoxy substituents on the benzene ring as shown. When the node size is 50, these 38 colchicine analogues are supplemented by 12 compounds which differ from them by having a six-member B-ring in place of the seven-member ring of the parent. These 12 additional compounds have an average activity of 7.31.

There are 17 colchicine analogues that were not found in any of the 10 trees listed in Table 6. The average activity of these 17 compounds, summarized in Figure 10b, is only 4.8. In each of these compounds, one or both of the 1,2-

Table 5. Summary of 10 RPSA Trees[a]

no. of trees	node size	mean activity	class size	class
10	54.50	7.43	50.50	anthracycline
10	52.90	7.09	52.00	camptothecin
10	73.50	6.34	69.40	podophyllotoxins
10	53.80	7.26	52.40	taxane
10	59.70	7.67	43.90	colchicine
10	53.20	6.65	19.70	macrolides
10	55.43	6.05	17.86	quinone, 2-amino
6	58.00	6.96	20.00	actinomycins
6	66.50	6.18	51.17	acridines, 9-amino
4	56.00	6.03	39.25	ellipticine
4	109.00	6.10	108.00	pyridine, 2-carbonyl(thiosemicarbazone)
averages	62.96	6.71	47.65	

[a] Parameters $K = 2$ features, minimum node size = 50. The column labeled *Node Size* is the average size of the tree node containing the class listed, *Class Size* is average number of class members in the node, and *Mean Activity* is the average pGI_{50} value. All statistics are averages over the number of tree nodes in which the class appeared.

Table 6. Distribution of Cancer Agents among Nodes in the RP Tree of Figure 4[a]

node	mean activity	node size	class size	class
21	5.89	127	104	pyridine, 2-carbonyl(thiosemicarbazone)
31	4.60	1065	104	acridines, 9-amino
			29	ellipticines
35	5.47	55	10	acridines, 9-amino
37	5.17	54	24	ellipticines
38	6.17	54	12	ellipticines
41	5.96	100	58	acridines, 9-amino
43	4.91	62	16	colchicines
44	7.50	72	62	colchicines
55	4.74	53	9	camptothecins
77	6.38	61	53	podophyllotoxins
78	6.67	65	28	camptothecins
			13	macrolides
80	6.38	99	18	camptothecins
			30	podophyllotoxins
81	7.42	76	17	actinomycins
			43	taxanes
94	7.08	76	63	anthracyclines

[a] The column labeled *Node* identifies the node in Figure 4 containing the class. The column labeled *Node Size* is the average size of the tree node containing the class listed, *Class Size* is average number of class members in the node, and *Mean Activity* is the average pGI_{50} value.

dimethyoxy substituents on the benzene ring was replaced by a hydroxy group. The remaining 23 colchicines, including the parent compound, all contain a 2-methoxy group on the tropone ring as illustrated in Figure 10c. This subclass has an average activity of 6.94, intermediate between the two thiomethyl tropone subclasses. Thus, the colchicines have very diverse activities, and RP/SA tends to find the most active subclass.

Several well-known classes of anticancer agents do not appear in the RP/SA trees shown in Figure 7 and summarized in Tables 3a,b and 5. These include alkylating agents, dihydrofolate reductase inhibitors such as methotrexate, and DNA and RNA synthesis inhibitors. Alkylating agents are structurally diverse, including nitrogen mustards, platinum compounds, and mitomycin, an aziridinyl indoledione. Methotrexate and its analogues are structurally homogeneous, but they comprise a small class of 16 compounds and have only moderate activity with an average value of 6.04 against the

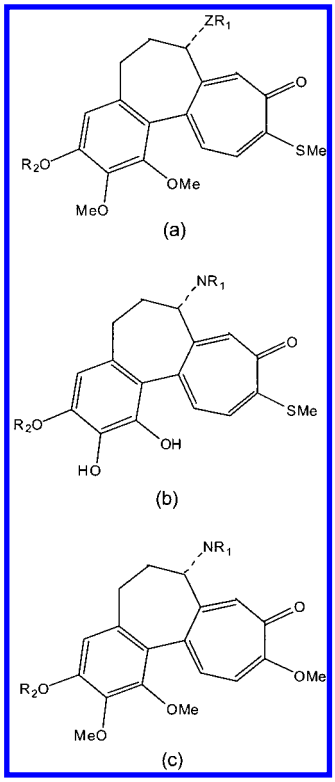

Figure 10. Colchicine subclasses.

NCI−H23 cell line. Similarly, other structurally homogeneous classes with moderate activity will appear at lower levels of the RP/SA tree, but these classes cannot compete with the classes identified in Table 5.

Comparing the trees generated by RP and RP/SA, Figures 4 and 7, respectively, reveals that the RP/SA trees have a much simpler structure. With our feature set, they are generally linear or close to linear, and at each split, the (+) branch is a terminal node consisting of a small group of compounds with high mean activity and set size approximately equal to the minimum frequency. Although RP/SA is a stochastic process that produces a different tree with each execution, with the NCI set it consistently identifies the most potent classes of anticancer agents. In addition, the established classes are cleanly separated from each other.

The results in Figure 5 show that combinations of two features produce groups of compounds with high mean activity. Using combinations of more features produces only a modest improvement. This behavior is very dependent on the feature set. Many LeadScope structural features, such as *benzene, 1-(2-aminoethyl)-*, *4-methoxy-*, are quite precise and characterize the active subclasses in the NCI quite well. With other molecular descriptors, such as atom-pairs,[4] combinations of greater numbers of features will probably be needed to achieve the same level of structural homogeneity.

We believe RP/SA will prove to be a valuable tool for identifying structurally homogeneous classes with high mean activity and also for discovering unexpected structure−activity relationships. Since RP/SA is a stochastic process, we believe it can be most effectively used as an exploratory tool if it is run multiple times, varying the control parameters, and the several results are cross compared. We are currently studying ways to merge the results from several parallel trees. Although it is possible to use large RP/SA trees to predict the activity of a new compound by averaging the predictions

from individual trees, we are investigating other refinements to RP/SA that appear more promising for this task. For the moment, we regard RP/SA as a valuable exploratory tool.

REFERENCES AND NOTES

(1) Hawkins, D. M.; Kass, G. V. In *Topics in Applied Multivariate Analysis*; Hawkins, D. H., Ed.; Cambridge University Press: Cambridge, U.K., 1982; pp 269−302.

(2) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees*; Wadsworth: New York, 1984.

(3) 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.

(4) 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.

(5) Jones-Hertzog, D. K.; Mukhopadhyay, P.; Keefer, C. E.; Young, S. S. Use of recursive partitioning in the sequential screening of G-protein-coupled receptors. *J. Pharmacol. Toxicol. Methods* **1999**, *42*, 207−215.

(6) Chen, X.; Rusinko, A.; 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.

(7) Chen, X.; Rusinko, A.; Tropsha, A.; Young, S. S. Automated pharmacophore identification for large chemical data sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 887−896.

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

(9) 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.

(10) Kirkpatrick, S.; Gelatt C. D.; Vecchi, M. P. Optimization by Simulated Annealing. *Science* **1983**, *220*, 67−68.

(11) Zheng, W.; Cho, S. J.; Waller, C. L.; Tropsha, A. Rational Combinatorial Library Design. 3. Simulated Annealing Guided Evaluation (SAGE) of Molecular Diversity: A Novel Computational Tool for Universal Library Design and Database Mining. *J. Chem Inf. Comput. Sci.* **1999**, *39*, 738−746.

(12) Holland, J. *Adaptation in Natural and Artificial Systems;* University of Michigan Press: 1975.

(13) Gobbi, A.; Poppinger, D.; Rohde, B. First International Electronic Conference on Synthetic Organic Chemistry; 1997, http://www.unibas.ch/mdpi/ecsoc/f0008/f0008.htm.

(14) Dorigo, M.; Gambardella, L. M. *IEEE Trans. Evol. Comput.* **1997**, *1*, 53−66.

(15) 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.

(16) Nonconfidential screening results and chemical structural data from the NCI's Developmental Therapeutics Program are available online: Human Tumor Cell Line Screen; http://dtp.nci.nih.gov/docs/cancer/cancer_data.html.

(17) Boyd, M. R.; Paull, K. D. Some Practical Consideration and Applications of the National Cancer Institute In Vitro Anticancer Drug Discovery Screen. *Drug Dev. Des.* **1995**, *34*, 91−109.

(18) Roberts, G.; Myatt, G. J.; Johnson, W. P.; Cross, K. P.; Blower, P. E., Jr. LeadScope: Software for Exploring Large Sets of Screening Data. *J. Chem. Inf. Comput. Sci.* **2000,** *40*, 1302−14.

(19) Kearsley, S. K.; Sallamack, S.; Fluder; E. M., Andose, J. D.; Mosley, R. T.; Sheridan, R. P. Chemical Similarity Using Physiochemical Property Descriptors. *J. Chem. Inf. Comput. Sci*. **1996**, *36*, 118−127.

(20) *Cancer Chemotherapeutic Agents*; Foye, W. E., Ed.; American Chemical Society: Washington, DC, 1993.

(21) See the anticancer maps produced by the Covell Group, National Cancer Institute: http://spheroid.ncifcrf.gov/anti_cancer.cfm.

(22) Lu, M. C. Antimitotic Agents. In *Cancer Chemotherapeutic Agents*; Foye, W. E., Ed.; American Chemical Society: Washington, DC, 1993; Chapter 9, pp 345−63.

CI0101049