Analysis of a Large Structure/Biological Activity Data Set Using Recursive Partitioning and Simulated Annealing

Student: Ke Zhang

MBMA Committee: Dr. Charles E. Smith (Chair)

Dr. Jacqueline M. Hughes-Oliver

Dr. Jason A. Osborne

Advisor for this project: Dr. Jacqueline M. Hughes-Oliver

02/2003

Project Report for MBMA - Institute of Statistics Mimeo Series #2542

Analysis of a Large Structure/Biological Activity Data Set Using Recursive Partitioning and Simulated Annealing

ABSTRACT

Large quantities of structure and biological activity data are quickly accumulated with the development of high-throughput screening (HTS) and combinatorial chemistry. Analysis of structure-activity relationships (SAR) from such large data sets is becoming challenging. Recursive partitioning (RP) is a statistical method that can identify SAR rules for classes of compounds that are acting through different mechanisms in the same data set. We use a newly proposed method called Recursive Partitioning and Simulated Annealing (RP/SA) to produce stochastic regression trees for biological activity. In the new algorithm a set of structural descriptors is extracted at each splitting node by using SA as a stochastic optimization tool. For one data set, results show that RP/SA is advantageous in analyzing the SAR information.

INTRODUCTION

With the development of high throughput screening (HTS) techniques[1,2], large structure/biological activity data sets are becoming available. Consequently, effective methods that can manage this amount of data and convert it into utilizable information are in great demand. It is well known that molecular structure is highly related to biological activity. Two chemical compounds with similar chemical structure are very likely to have similar biological potency with respect to one or more assays. Therefore, forming Structure Activity Relationships (SAR) plays a critical role in analyzing chemical data sets. The task is becoming more challenging when the size of data set

increases. The simple fact of the matter is that there may simultaneously exist more than one SAR rule in a large data set such that traditional statistical methods are no longer suitable to determine them.

In this paper, we use a newly proposed method, RP/SA (Recursive Partitioning and Simulated Annealing) [3,4] to analyze a new data set. The method shows several advantages in mining large and diverse chemical data sets from high throughput screening. The purpose of RP/SA is to divide a large heterogeneous compound set into several more homogenous subsets based on different SAR rules. RP/SA uses a combination of descriptors for each split in a recursive partitioning algorithm, which is globally optimized by utilizing a stochastic search. The main distinction between RP/SA and standard RP or other methods is that the searching space of RP/SA for optimization includes all possible combinations of descriptors at each step, whereas standard RP is trying to search for an optimal split in a space consisting of only individual descriptors. It is reported 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 (Blower et al., 2002, p.393). " Therefore, RP/SA can be expected to perform better than standard RP. The main problem existing in choosing the best combination of descriptors for a split is computational. Suppose there are 1000 descriptors, the tentative searching space may have 100 million combinations of three descriptors. A step-by-step global search for the best set of three descriptors requires a comparison among over 100 million possible splits. The situation is worse when the total number of descriptors increases or the number of descriptors in a splitting set increases.

One possible solution to this is to use a stochastic search algorithm. Simulated Annealing (SA)[5] is one of the most popular methods. The concept is based on the manner in which liquids freeze or metals re-crystalize in the process of annealing. SA is a global optimization method that distinguishes between different local optima. When maximizing the splitting criterion in our algorithm, SA takes a step from an initial value then evaluates the function. It accepts all uphill steps and some downhill steps such that it is able to avoid becoming trapped at a local optimum. Due to the stochastic nature of the algorithm, resulting RP trees will be different with each run. However, it can be expected that several different sets of descriptors may identify the same class of compounds.

METHODS

Database. The database used in this study is obtained from an experiment conducted by GlaxoSmithKline.[6] "1000 chemically diverse compounds were selected from U.S. liquid stores. The selection is based on Burden numbers. Compounds with similar Burden numbers are more likely to be topologically similar. A systematic sample of 1000 is chosen from U.S. liquid store compounds ordered by their Burden numbers. Compounds are screened for biological potency with respect to an assay. The biological data, percent inhibition, is continuous. In theory, the response should be in the range 1-100, but in practice, results can be outside this range. Percent inhibition is calculated according to an equation on the target compound and the reference compound, which can be less than 0 or greater than 100. Certain binding situations, e.g., a compound binding to a secondary binding site, can lead to unusual results, thus yielding a percent inhibition outside the typical range. But because percent inhibition

is discretized into potent or not according to specified thresholds, the extreme values below 0 and above 100 do not cause a problem. The cutoff thresholds are 60. Of the 1000 compounds tested, only 40 (4%) are potent for the assay (Zhu et al., 2001, p.924)."

Atom Pair Descriptor. Atom pair descriptors[7, 8] are composed of two components ---atom types and the minimal distance evaluated by the number of atoms in the shortest path between them. From the conceptual standpoint, the term "atom" here consists of the key chemical or physical feature points and the spatial relationships between them. Therefore, an atom can represent a real atom in the chemical structure, such as hydrogen, carbon, or oxygen, or the center of some special chemical functionality, such as aromatic ring, hydrogen bond donor, positive charge center, negative charge center, etc. Each atom pair should include two key features and the spatial relationship between them. Consequently, there exists *m(m-1)/2* atom pairs (where *m* is the number of non-hydrogen atoms in a structure) in each structure. In our algorithm, all atom pair descriptors are binary variables with 1 indicating the presence and 0 indicating absence of a particular atom pair. 1873 descriptors are used in this study.

Splitting Criterion. The splitting criterion can be difference in mean biological potency of the two groups, or the *F*-statistic for comparing two groups (which is equivalent to *t*-statistics for comparing two groups). The standard splitting criterion used by RP/SA is characterized by

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

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 subsets, respectively.

If $n_+ << n_-$, which is commonly the case when only a very small fraction of the compounds in a collection is potent enough to act as lead molecules in later drug discovery phases, this t-criterion can be approximated by

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

where σ is the standard deviation of the node being split. This usually occurs in the early stage of the tree. Thus, the t criterion tends to give the largest value of $\sqrt{n_+}D$ rather than the largest absolute difference.

An alternative splitting criterion is absolute difference between the two means (i.e. *D*-value). It is critical that the *D* criterion works with appropriate minimum node size. This is very dependent on the number of compounds.

Simulated Annealing Portion of RP/SA. 1. This procedure is shown as a flowchart in Figure 1. Some initial conditions should be specified first such as minimum node size *n*, number of descriptors used in splitting set *K*, an initial temperature T_0, a minimum temperature cutoff T_{min}, and a temperature reduction rate α. In addition, a descriptor pool from which a combination of *K* descriptors will be chosen randomly should be initialized as well. The descriptor pool is obtained by selecting only those descriptors that exist in at least *n* compounds.

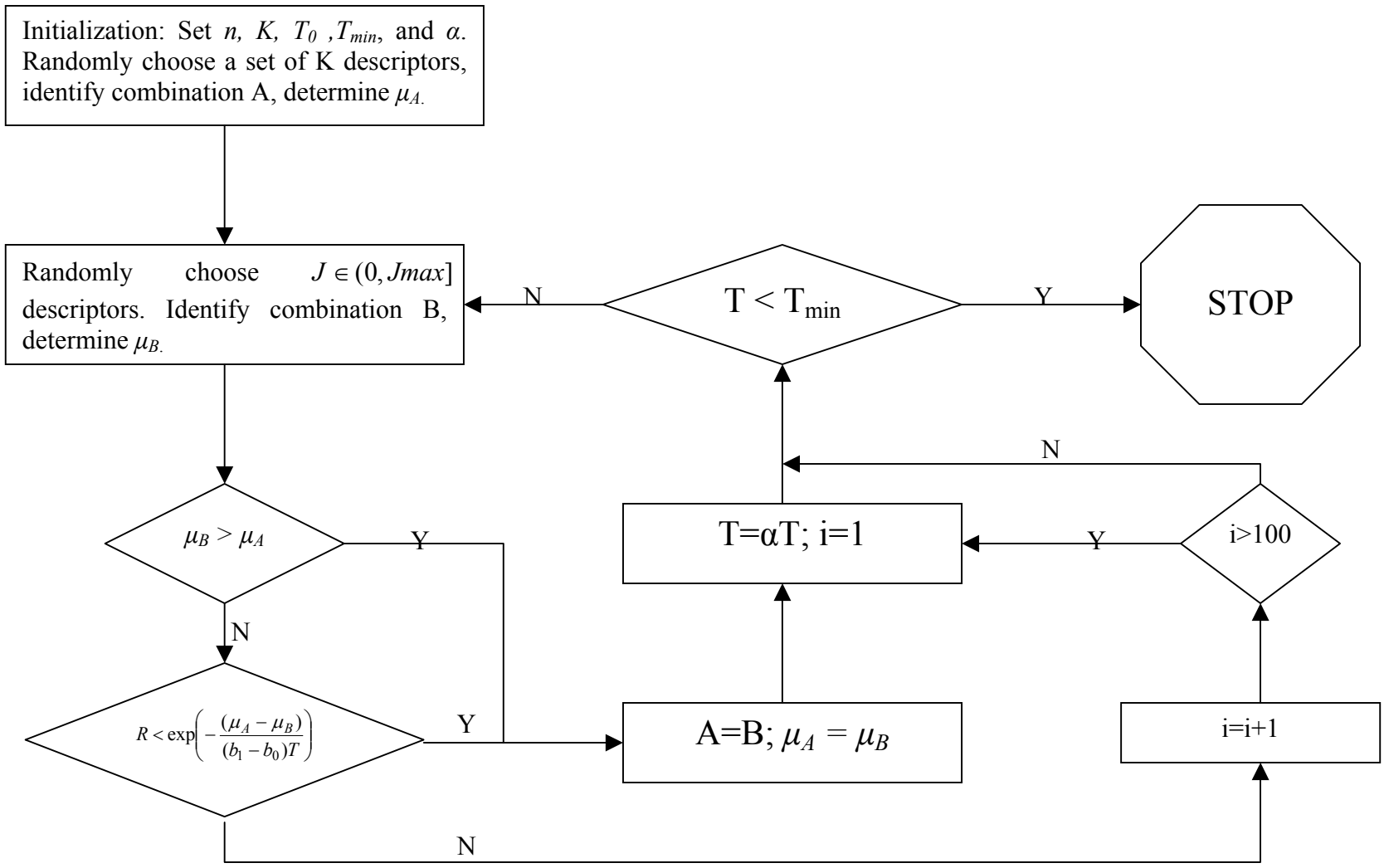

Figure 1. Simulated annealing algorithm. *A* and *B* denotes subsets of *K* molecular descriptors and also the corresponding compound sets; *i* is a counter which records the number of iterations since *T* was lowered; b_0 is the mean activity of the full set; and b_1 is the threshold activity for determining active compounds. J_{max} decreases with *T* according to the schedule described in the text. (Adapted from Blower et al., 2002, p.396)

2. A combination of K descriptors is randomly chosen from the descriptor pool. The number of compounds that simultaneously contain this K-descriptor combination is determined. If it is less than the minimum node size n, another combination of K descriptors will be randomly selected until the number is larger than or equal to n. The satisfying combination of K descriptors results in the set A of compounds representing the positive subset (a subset of compounds containing all K descriptors). A corresponding splitting value μ_A is calculated according to the splitting criterion.

3. Random number $J \in (0, J_{\max}]$ is chosen. The upper bound J_{max} is a function of the temperature T, which has the following form

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

J randomly selected descriptors are dropped from the current combination of K descriptors. J randomly selected descriptors chosen from the descriptor pool replace the J dropped descriptors. It is obvious that J_{max} gets smaller as T decreases. Consequently, the value of J is small when temperature becomes low. Thus the combination of K descriptors becomes more stable in the low temperature. The J descriptors are chosen repeatedly until the minimum node size n requirement is satisfied. With a new set of K descriptors, the resulting positive subset B of compounds that contain all of the K newly chosen descriptors is determined. We calculate the value μ_B of the splitting criterion.

4. If $\mu_B > \mu_A$, the new set of K descriptors is accepted as a better combination and used as the current set of K descriptors. The temperature T is decreased, $T_{new} = \alpha T$ where $\alpha \in (0, 1)$ and the procedure returns to step (3).

If $\mu_B \leq \mu_A$, a uniform random number R is chosen from [0,1). A check is made of the Metropolis condition to determine if the new set of K descriptors can be accepted:

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

where b_0 is the mean activity of the full set and b_1 is the threshold activity for determining active compounds. If the above inequality is satisfied, the K descriptors with the new subset of J descriptors are used as current set of K descriptors. The temperature T is lowered to $T_{new} = \alpha T$ where $\alpha \in (0, 1)$ and the procedure returns to step (3).

5. If the two criteria in (4) are not satisfied, the new K descriptors are rejected. Steps (3)-(5) are repeated until one of the two criteria in (4) is met. If the loop procedure fails to satisfy the two criteria for 100 times, the temperature T is lowered to $T_{new} = \alpha T$ where $\alpha \in (0, 1)$ and the procedure returns to step (3).

6. The Simulated Annealing process terminates when the current temperature is less than the minimum temperature T_{min} which is specified at the beginning. Consequently, the current K descriptors are used to divide a node into two subsets and subset A represents the positive (+) response group.

Recursive Partitioning Portion of RP/SA. The RP is to use SA algorithm recursively in splitting nodes until the stopping criteria described below are met. Thus, a stochastic regression tree can be obtained in this step.

Stopping Criteria. There are several methods that can stop the tree from growing. One simple and straightforward way is to specify a maximum depth to which to grow the tree. The value of maximum depth depends on practical limitations of time and complexity. Maximum depth is acting as one of three stopping criteria in current version of RP/SA. The other two are: (i) A node will be a terminal node if the maximum potency of individual compounds in the node is less than b_1; (ii) A node will be a terminal node if the node size is less than or equal to n.

RESULTS AND DISCUSSION

Standard RP/SA Trees. When using SA on our dataset, input parameters to the SA algorithm were set, following Blower et al, as T_0=10, T_{min}=10^{-3}, and α=0.9. Minimum node size is set at n=5, K=3 descriptors are considered and potency threshold is b_1=60, with an observed mean activity of b_0=7.9. Since RP/SA has a stochastic component in the algorithm it produces a different tree with each run. One RP/SA tree is shown in Figure 2. This figure is read in the following way. In the topmost (root) node there are 1000 compounds with an average potency of 7.9. The variability of these compounds is measured with the standard deviation of the individual compound. The "best" set of K=3 descriptors for splitting, as determined by stochastic search, is written below the node. Compounds simultaneously containing all of the "best" K descriptors are split to the right; others are split to the left. The rules tracing to a terminal node give the descriptors that define a class of compounds. This tree is grown using three descriptors, maximized t-criterion for splitting and a minimum node size of 5. There are two active nodes in this RP/SA tree, whose average potencies are greater than 60. These nodes are marked by colored effect. One of these nodes contains 1 compound

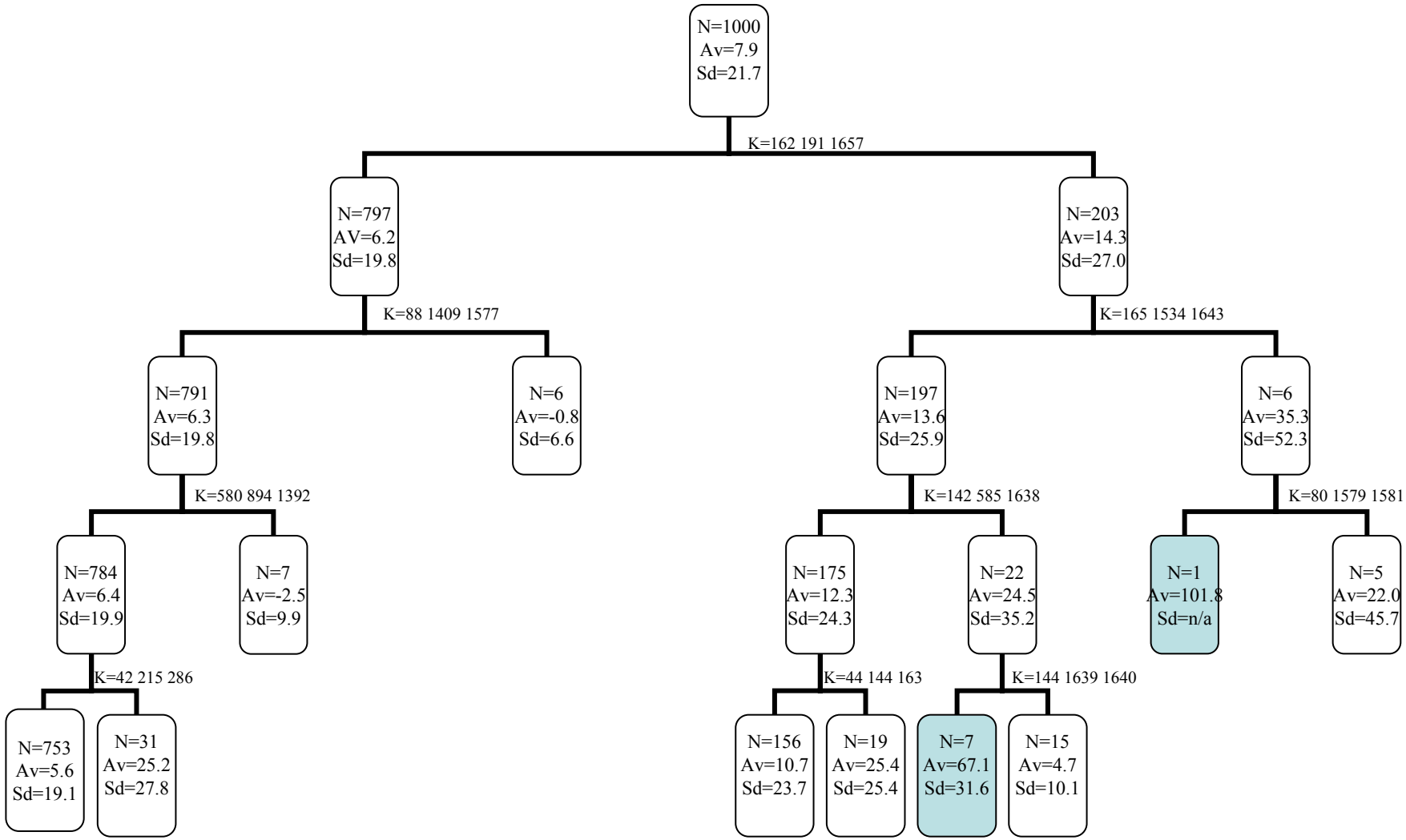

Figure 2: A sample RP/SA tree; parameters: splitting criteria = t-critirion, K = 3 features in combination, minimum node size n =5. For each node, "N" is the number of compounds, "Av" represents the average potency of the compounds, "Sd" represents the standard deviation of the potencies of the compounds. The potent nodes are indicated as the colored nodes.

and the other contains 7 compounds. Since these compounds are in terminal nodes that have very different key descriptor combinations, it is very likely that they indicate the existence of several mechanisms of activity in this data set.

To quantify the effectiveness of a tree in identifying active groups of compounds, we can use two methods. A simple and straightforward way is to calculate the average activity of potent nodes in the tree. The other method is to use the average activity of "better" daughter nodes (the node with higher average potency is a better node for each split). More precisely, we first specify the depth *d* at which the trees to be compared terminate. Let a_1 be the average activity of all potent nodes occurring above or at depth d; let a_2 be the average activity of remaining "better" daughter nodes (including non-terminal nodes but not including potent nodes); and let *c* be the number of all potent nodes occurring above or at depth *d*. Then our measure of "quickly found" activity is given by

$$M= (c/2^{(d-1)})a_1 + (1- c/2^{(d-1)})a_2,$$

with larger values of *M* being desirable. It is easy to show the maximum value of *c* is $2^{(d-1)}$ according to our stopping criteria. Thus, the *M* method is a weighted average between a_1 and a_2, which is different from *Q* method Blower et al proposed in 2002. The average activity of positive terminal nodes is the only factor that is taken into account in *Q* measure, whereas both average activity of positive terminal (potent) nodes and that of remaining "better" daughter nodes are counted into *M* measure. It is possible that two trees with the same number of positive terminal nodes but large difference between the depth yield similar values of *Q* by ignoring the effect of all non-terminal "better" daughter nodes. Therefore, the M measure can be expected to give a fairer comparison because of considering more useful information. It also has the advantage

of not having to grow the tree too far for comparison purpose. The value of d varies according to different dataset. We used d=4 for our dataset. For Figure 2, there are c=2 potent nodes with average activity a_1 =84.5. There are seven remaining "better" daughter nodes with average activity a_2=19.6. This yields M=35.8.

RP/SA Trees Using Mean Difference Criterion. A simple adjustment to the standard RP tree is to replace the t-criterion for splitting a node with the simpler criterion of the difference in means. For our database, this leads to a remarkable improvement in detecting the more active groups of compounds. A tree using three descriptors, the same stochastic generating seed, minimum node size of 5 and mean difference criterion is shown as Figure 3. A visual comparison of Figure 3 with Figure 2 shows that the number of potent nodes and the biological activity of "better" daughter nodes found by RP/SA with mean difference criterion are larger than those found by the standard RP/SA. It is shown in Table 1 that the M values for this procedure range about 100% higher than those for standard RP/SA. Moreover, as seen in Figure 3, the tree resulting from RP/SA using the mean criterion is quite a bit simpler than the tree from the standard RP/SA.

Table 1. M-value Comparison between Standard RP/SA and RP/SA Using Mean Difference Criterion

	Standard RP/SA	RP/SA Using Mean Difference Criterion
1. (K=3, n=5, α=0.90)	35.8	59.6
2. (K=3, n=5, α=0.90)	24.6	53.6
3. (K=3, n=5, α=0.90)	20.1	63.8
4. (K=3, n=5, α=0.99)	26.6	63.0
5. (K=4, n=5, α=0.90)	29.6	56.6

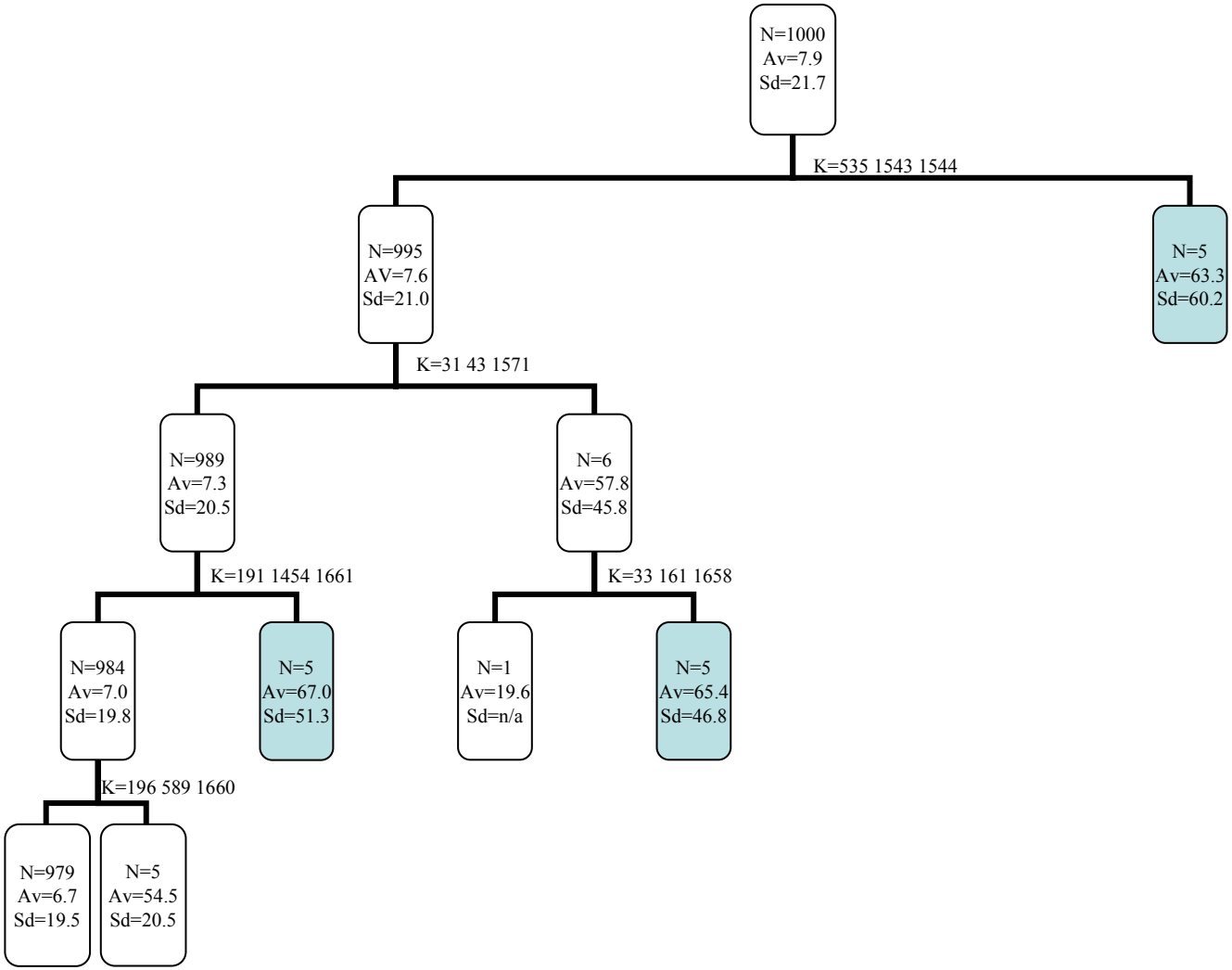

Figure 3: A sample RP/SA tree; parameters: splitting criteria = absolute mean difference, *K = 3* features in combination, minimum node size *n* =5. mean difference. For each node, "N" is the number of compounds, "Av" represents the average potency of the compounds, "Sd" represents the standard deviation of the potencies of the compounds. The potent nodes are indicated as the colored nodes.

Because of the stochastic component in the algorithm, we generated five trees using the same conditions that were used to create Figure 3. Actually, the average activity of the "better" daughter nodes (M value) was moderately stable over the 5 trees produced. Figure 4 illustrates the trees corresponding the best and the worst M values over the 5 trees. The figure indicates the lowest and highest *M* values are 53.6, 63.8, respectively. In addition, the randomness is beneficial in exploring different solutions and determining which parts of the solutions are stable.

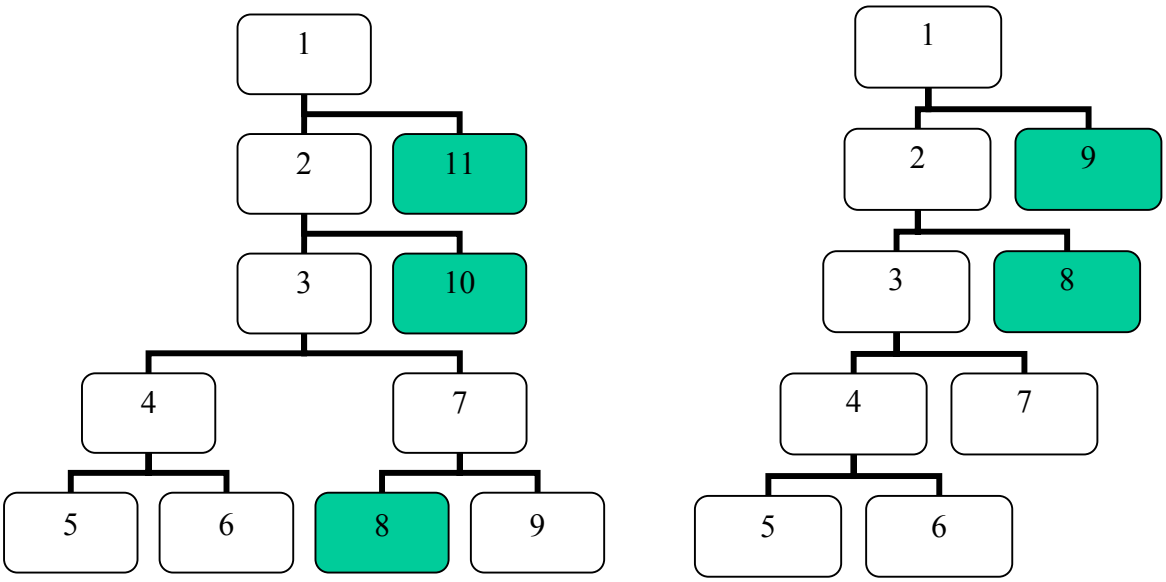

Figure 4: Sample RP/SA trees using mean difference criterion. RP/SA parameters: *K*=3 descriptors and minimum node size = 5. Potent nodes (average potency > 60) are highlighted. Tree 1 (left) – M is 63.8; tree 2 (right) – M is 53.6.

For our descriptor set and data, a noticeable improvement was gained by using three descriptors as opposed to two. However, using a combination of more descriptors produces little improvement. This behavior is very dependent on the descriptor set. In other words, this may not occur for descriptor sets that have more prevalent descriptors, since in that case more descriptors are necessary to identify potent nodes.

Because of the random effect of the SA portion of our algorithm, we wondered about the need to conduct a more thorough search. To investigate this, we modified two main stochastic steps in RP/SA. First, the check for a Metropolis condition was deleted from the algorithm. Thus, the new set of *K* descriptors would replace the old one only if $\mu_B > \mu_A$. RP/SA parameters were the same as those used above in RP/SA using mean difference criterion. The *M* value was 52.0, illustrating that deletion of the Metropolis condition does not necessarily lead to a better solution. In hindsight, the Metropolis condition decreases the possibility of convergence to a local maximum, so removing it probably led to a sub-optimal splitting set. Moreover, the Metropolis condition can reduce run times, especially when a more complicated splitting criterion is chosen.

The other adjustment was to grow an "almost-exhaustive" tree, where at each level we ran simulated annealing 10 times (we set α equal to 0.99). The M value for this tree was 63.0 and was comparable to the values found by the regular RP/SA trees.

Although RP/SA is designed to identify groups of active compounds having a few key descriptors in common, RP/SA tends to find active compounds having many additional common things such as the absence of several descriptors in common. This was supported by observation of potent nodes in negative branches. Its benefits include finding unexpected structure-activity relationships. However, the interpretation of some potent nodes can be difficult because negative nodes cannot tell you exactly which descriptors of the set are absent and what are present in compounds. One possible solution is to consider an extended searching space. Suppose a set of three descriptors would be chosen to split the node. The presence of two descriptors

and the absence of one descriptor should be taken into account as well as the simultaneous presence of three descriptors. The corresponding positive branch would consist of compounds that contain two of three descriptors and miss the other one. Consequently, there are four possible cases for any given three descriptors ($\binom{3}{2}+\binom{3}{3}=4$). This adjustment results in a longer computational time and requires a more effective stochastic searching method that can guarantee a high probability of finding a comparable result within a reasonable time.

SUMMARY

From the results of this work, it is demonstrated that RP/SA is a valuable tool for identifying different structure-activity relationship rules in a very large data set. Since RP/SA is a stochastic algorithm, we believe that under multiple runs it can identify different subsets of descriptors that may classify the same group of compounds. It is a promising work to produce a rational score scale from stochastic trees to predict biological activity of new compounds.

REFERENCE

[1] Sittampalam, G. S.; Kahl, S. D.; Janzen, W. P. High-throughput screening: advances in assay technologies. *Curr. Opin. Chem. Biol.* **1997**, 1, 384-391.

[2] Silverman, L; Campbell, R.; Broach, J. R. New assay technologies for high-throughput screening. *Curr. Opin. Chem. Biol.* **1998**, 2, 397-403.

[3] 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.

[4] Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees;* Wadsworth: New York, 1984.

[5] Kirkpatrick, S.; Gelatt, C. D.; Vecchi, M. P. Optimization by Simulated Annealing. *Science* **1983**, 220, 671-680.

[6] Zhu, L. Hughes-Oliver, J. M.; Young, S. S. Statistical Decoding of Potent Pools Based on Chemical Structure. *Biometrics.* **2001**, 57, 922-930.

[7] Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, 25, 64-73

[8] 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.