Analysis of a Large Structure/Biological Activity Data Set Using Recursive Partitioning[1]

Andrew Rusinko, III,[†] Mark W. Farmen,[‡] Christophe G. Lambert,[§] Paul L. Brown, and S. Stanley Young*

Research Information Systems, Glaxo Wellcome Inc., Five Moore Drive, Research Triangle Park, North Carolina 27709

Received February 25, 1999

Combinatorial chemistry and high-throughput screening are revolutionizing the process of lead discovery in the pharmaceutical industry. Large numbers of structures and vast quantities of biological assay data are quickly being accumulated, overwhelming traditional structure/activity relationship (SAR) analysis technologies. Recursive partitioning is a method for statistically determining rules that classify objects into similar categories or, in this case, structures into groups of molecules with similar potencies. SCAM is a computer program implemented to make extremely efficient use of this methodology. Depending on the size of the data set, rules explaining biological data can be determined interactively. An example data set of 1650 monoamine oxidase inhibitors exemplifies the method, yielding substructural rules and leading to general classifications of these inhibitors. The method scales linearly with the number of descriptors, so hundreds of thousands of structures can be analyzed utilizing thousands to millions of molecular descriptors. There are currently no methods to deal with statistical analysis problems of this size. An important aspect of this analysis is the ability to deal with mixtures, i.e., identify SAR rules for classes of compounds in the same data set that might be binding in different ways. Most current quantitative structure/activity relationship methods require that the compounds follow a single mechanism. Advantages and limitations of this methodology are presented.

INTRODUCTION

Combinatorial chemistry and high-throughput screening (HTS) are having a profound impact on the way pharmaceutical companies identify new therapeutic leads. Five to ten years ago, it was impressive to screen 50 000 compounds looking for a lead molecule. Now 50 000 compounds are routinely screened in a matter of weeks. Each compound can be described with tens to thousands of variables, so vast quantities of data are now routinely being produced from high-throughput biochemical assays. Furthermore, the construction of chemical libraries has effectively replaced the painstaking individual synthesis of compounds with a parallel synthesis of many thousands of compounds about a common scaffold.[2] Since there is a very low probability of identifying a new lead compound from any single compound tested in a screening program, it is expected that by testing vast numbers of compounds from inventory or made via a combinatorial approach, a sufficient number of novel leads will be found. However, the synthesis and testing of many thousands of compounds has placed a tremendous strain on the logistical and computational infrastructure relied upon to store and analyze these data sets. Methods developed and used in the past decade for the statistical analysis of a relatively small number of compounds (usually less than 100), e.g., COMFA, Catalyst, and MCASE, are not suitable for use on these much larger data sets. This is particularly true if the active compounds follow more than one mechanism. Consequently, new analysis techniques must be found and investigated. There is no published method for the analysis of hundreds of thousands of compounds each described with tens of thousands to millions of descriptors. We have such a method, and that is the subject of this paper.

Various methods for the storage and retrieval of structure−activity data have been devised in the past few years. GLIB[3] was developed by Glaxo Wellcome to address storage and retrieval of chemical library data. In addition, software products are now available from major vendors that address most of the logistical needs of combinatorial chemistry.[4] Unfortunately, there is little literature guidance on how the vast data might best be used to guide future screening and synthetic efforts. One analysis strategy is a simple sort of relevant biological data and selecting the most potent compounds for retesting to give a short list of promising new lead compounds for further consideration. Many research programs stop here and immediately revert to traditional methods to optimize the new leads. Typically, ligand−receptor docking[5] or three-dimensional (3D) pharmacophoric identification[6] studies are performed. Other than the receptor structure, often not available, the ligand docking studies make very little or no use of the screening data. Virtually all algorithms designed to identify three-dimensional pharmacophores cannot handle very large data sets or multiple binding modes/sites.

For a number of years, there has been an interest in using artificial intelligence or machine learning to uncover hidden rules from or otherwise classify chemical data sets.[7] Many have focused on reaction[8] or spectral[9] prediction. Others have used neural networks,[10] fuzzy adaptive least squares,[11] inductive logic,[12] machine learning,[13] or frequency-based

[†] Alcon Laboratories, Inc., 6210 S. Freeway, Fort Worth, TX 76134
[‡] Eli Lily and Co., Lilly Corporate Center, Indianapolis, IN 46285.
[§] Golden Helix Datamining, P.O. Box 10633, Bozeman, MT 59719.

10.1021/ci9903049 CCC: $18.00
Published on Web 10/30/1999

fragment weighting schemes[14] to analyze structure−activity data sets or predict chemical properties. "Data mining" [15] has become a fashionable buzzword with a World Wide Web site[16] devoted to methods for extracting information from large databases. These methods are often specialized for particular subject areas and are not designed for routine structure/activity relationship (SAR) analysis.

Alternatively, some researchers have employed a sequential approach to make best use of all the data. In essence, information gained from each set of screening experiments guides the next generation of experiments. Weber et al.[17] and Singh et al.[18] used genetic algorithms to select subsequent polypeptide libraries to be made. Gobbi et al.[19] used a genetic algorithm to select active compounds in a large database by iterative selection and screening cycles. Similarly, CombiChem[20] uses an initial set of compounds to generate a hypothesis, and then subsequent compound sets are chosen for synthesis and testing from a virtual library. While results appear encouraging, the latter two techniques are proprietary and have only been applied retrospectively to literature data sets.

Recursive partitioning (RP) is a simple, yet powerful, statistical method that seeks to uncover relationships in large complex data sets. These relationships may involve thresholds, interactions, and nonlinearities. Any or all of these factors impede an analysis that is based on assumptions of linearity such as multiple linear regression (or linear free energy relationships), principal component analysis (PCA), or partial least squares (PLS). RP was used to analyze a peptide library[21] data set and a set of monoamine oxidase (MAO) inhibitors using molecular fragment codes as descriptors.[22] Various implementations of RP exist, such as FIRM,[23] CART,[24] or C4.5,[25] but none have been specialized and extended to the specific problem of generating SAR from very large structure/activity data sets. In this paper, we describe a new computer program, SCAM (Statistical Classification of the Activities of Molecules), which can be used to analyze large numbers of *binary* descriptors (i.e., the presence or absence of a particular feature) and interactively partition a data set into activity classes. This program is unprecedented in the size of data sets it can handle. On an internal project we analyzed a data set of 300 000 compounds each described with over 2 000 000 descriptors. The analysis was completed in less than 1 h of computer time. It is quite possible to analyze even larger data sets. To illustrate this technique, we will use SCAM to uncover substructural rules that govern the biological activity of a publicly available set of 1650 MAO inhibitors[26] from Abbott Laboratories. Using these rules, it is possible to mathematically screen large collections of untested compounds for those molecules likely to be active, to design combinatorial libraries, and to guide the synthesis of individual compounds.

METHODS

Descriptors. To illustrate the method, three types of descriptors are used in our studies. Each had its origin in the pioneering work on molecular topological similarity at Lederle. Atom pairs,[14c] topological torsions,[27] and atom triples are descriptors generated from the topological (2D) representation of a molecular structure. Atom pairs are very simple descriptors composed of atoms separated by the minimal topological distance (i.e., the number of atoms in the shortest path) between them. As shown in Figure 1a, each local atomic environment is characterized by three values: the atomic number, the number of non-hydrogen connections, and half of all the associated π-electrons. For example, the terminal triple bond carbon in pargyline is encoded as C(1,2) while the connecting carbon of the phenyl ring is encoded as C(3,1). Thus, for each structure, $n(n-1)/2$ atom pairs (where n is the number of non-hydrogen atoms in a structure) are generated by considering each atom and the minimal topological distance to every other atom in turn. For example, the C(1,2) to C(3,1) path contains six atoms. For our purposes, a bit-string indicating the presence or absence of a particular atom pair is produced. From a data set of about 1000 drug-like structures, approximately 10 000 unique types of atom pairs are generated.

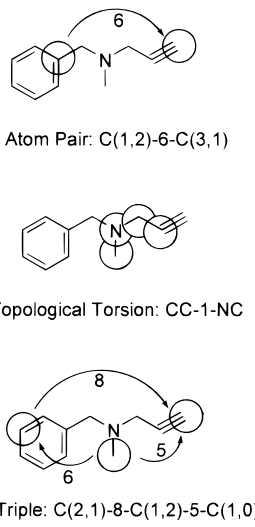


Figure 1. Example of the descriptors used in these calculations: atom pair, topological torsion, and atom triple. Atom code C(1,2) indicates a carbon atom with one connection and two π-electrons, i.e., sp-hybridized carbon. The total path length between each atom in a pair is also indicated.

Topological torsions[27] are also quite simple to generate. Originally, they were proposed to complement atom pairs since it was felt that atom pairs did not adequately capture a local substructural environment. Topological torsions are generated in the following manner. For each nonterminal bond, first note the atom types (as defined for atom pairs) that participate in the bond. Next, for each possible combination of atoms directly connected to those in the bond, output a topological torsion descriptor such as shown in Figure 1b. We modify the original definition in two ways. First, we include terminal bonds but reduce the size of the topological torsion descriptor to three atoms from four since there are obviously no atoms connected to the terminal atom. Second, the bond type is also explicitly recorded in the topological torsion. Both modifications have the effect of increasing the number of unique features encoding structural information.

The last structural descriptor type, an atom triple, is used by several groups for molecular similarity searching[28] and as search keys for 3D search[29] and docking[30] studies. A triple of atoms and the corresponding interatomic distances are thought to be the most elementary pharmacophore. Our atom triples differ from those previously defined. We replace the interatomic distance with the length of the shortest path

between each pair of atoms forming the triple. For example, as shown in Figure 1c, the triple formed among the terminal triple bond carbon, an aromatic carbon, and the terminal methyl of pargyline is C(1,2)−8−C(2,1)−6−C(1,0)−5−. All possible triples are generated, and each is canonicalized to a unique form. A bit-string noting the presence and absence of atom pairs, topological torsions, and/or atom triples is then formed. Depending upon the size and diversity of the data set, it is possible to generate hundreds of thousands to millions of unique atom triples.

Implementation Details. SCAM builds a classification tree using recursive partitioning of a large data set, on the order of hundreds of thousands of structures (and associated potencies), and an extremely large number of descriptor variables (tens of thousands to millions). RP was originally designed for automatic interaction detection by Morgan and Sonquest[31] and Hawkins and Kass.[23] SCAM achieves amazing speed partially through the use of binary variables. In addition to the recursive partitioning algorithms, a structure viewer serves as a bridge between the statistical analysis and the actual chemical structures to facilitate understanding of the analysis results. Structural features that the algorithm identifies as associated with activity are highlighted in the structure display. Furthermore, several advanced algorithmic modifications to traditional RP are necessary to produce a real-time analysis of such large data sets. Specifically, sparse matrix techniques are employed to store and manipulate the data. Pseudocode is given for the current implementation of binary recursive partitioning in Appendix A of the Supporting Information.

Three input files are needed for a SCAM analysis. A data file contains the compound names and potencies, a descriptor dictionary file contains a contextual decoding of each descriptor variable, and a binary file contains a record for each structure which indicates the presence of each descriptor in the structure (absence of a feature is imputed). For each descriptor, a list of compounds containing the descriptor is formed. This conserves memory and is very similar to the concept of indirect keys used in substructure search. The alternative would be to store a complete list of descriptors for each structure. However, for our purposes the former is more efficient, since a *t*-test is performed on the activities of the structures associated with a particular descriptor.

In contrast to data partitioning via continuous variables, binary classification trees can be computed very quickly and efficiently since there are far fewer and much simpler computations involved. For example, FIRM[23] develops rules for splitting based on "binning" of continuous variables and amalgamating contiguous groups. These processes add considerably to execution time and hence limit the interactive nature of most general recursive partitioning packages to data sets much smaller than those under consideration. With binary data, on the other hand, a parent node can only be split into two and only two daughter nodes. Splitting of a binary descriptor such as the presence or absence of an atom pair involves performing a *t*-test between the mean of the group that has that feature and the group that does not. The best variable for a potential split can then be selected by using the largest *t*-statistic. Therefore, the *p*-value (a potentially time-consuming part of the calculation) needs only to be computed for the most significant split. Frequently, either the group that has the atom pair or the group that does not have the atom pair is quite small. This fact can be exploited using an idea known as "updating". Updating involves computing the necessary statistics in the smaller daughter node along with the statistics of the parent node to compute the statistics for the other daughter without reprocessing the entire collection.

If one denotes the potencies in group 1 by $x_1, x_2, ..., x_m$ and group 2 by $y_1, y_2, ..., y_n$ and assuming that group 1 is smaller than group 2 ($m < n$), the *t*-statistic for testing for a difference between group potency means is

$$t = \frac{[(\bar{x} - \bar{y})/(1/m + 1/n)]^{1/2}}{[(\mathrm{SSX} + \mathrm{SSY})/(n + m - 2)]^{1/2}} \quad (1)$$

where

$$\mathrm{SSX} = \sum_{i=1}^{m}(x_i - \bar{x})^2, \quad \bar{x} = \mathrm{SX}/m, \quad \mathrm{SX} = \sum_{i=1}^{m} x_i \quad (2)$$

$$\mathrm{SSY} = \sum_{i=1}^{n}(y_i - \bar{y})^2, \quad \bar{y} = \mathrm{SY}/n, \quad \mathrm{SY} = \sum_{i=1}^{n} y_i \quad (3)$$

Next, let $z_1, z_2, ..., z_{m+n}$ denote the potencies in the parent node. The sum, SZ, was computed for the previous split, so it is available. (Equation 1 is a simple reformulation of the equation for the *t*-statistic in Appendix A of the Supporting Information.) Therefore, after SX is computed, SY can be computed as the difference SY = SZ − SX.

A similar updating method can be used to compute SSX and SSY. Note that

$$\mathrm{SSX} = \sum_{i=1}^{m} x_i^2 - n\bar{x}^2 \quad (4)$$

$$\mathrm{SSY} = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 \quad (5)$$

so SSY can be computed using the sum of the data, SY, and the sum of the squared data which will be denoted by SYY. Having computed SXX, and having SZZ available, SYY can be computed by the relation SYY = SZZ − SXX. Therefore, the *t*-statistic can be computed very quickly, having stored the sum of the data and the sum of the squared data from the previous split. SCAM runs in time proportional to the size of the sparse matrix, and since for each descriptor the calculations run only over the structures with the feature, SCAM scales nearly linearly with the total number of descriptors.

RESULTS

Example Analyses. A small set of 1650 compounds tested as monoamine oxidase (MAO) inhibitors[26] is used to illustrate the effectiveness of SCAM in the analysis of large structure/activity data sets and producing SAR rules. Neuronal monoamine oxidase [amine:oxygen oxidoreductase (deaminating) E.C. 1.4.3.4] inactivates neurotransmitters such as norepinephrine by converting the amino group to an aldehyde (Scheme 1). Inhibitors of this enzyme were considered for the treatment of depression and were introduced into therapy in 1957 with the drug pargyline (Chart 1). However, due to toxicity and interactions with other drugs and food, MAO

Scheme 1

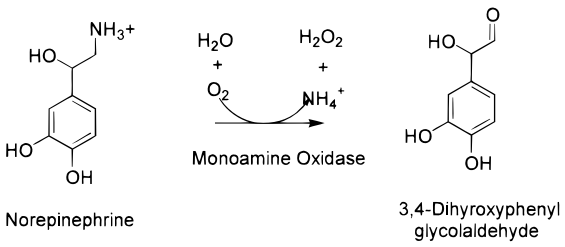
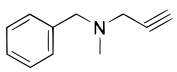


Chart 1



inhibitors are now rarely prescribed, yet reviews[32] touch upon the continued interest by the pharmaceutical community in MAO as a target for rational drug design in antidepressant therapy. Biological activities for this data set are reported in four classes: 0 being inactive, 1 somewhat active, 2 modestly active, and 3 most active. Generating a quantitative structure/activity relationship (QSAR) from this data set is problematic since the potencies are not precisely given and the heterogeneous nature of the data set implies that the compounds do not necessarily follow the same biological mechanism.

Recursive partitioning was applied to this set of 1650 activities using 6405 unique atom pairs and topological torsions as descriptors. The first five split-levels of the resulting tree diagram are shown in Figure 2. Default settings for SCAM were used to produce this tree. These are as follows: allow up to 10 levels of partitioning; only split on statistically significant variables (Bonferroni-adjusted p-value < 0.01); allow both positive and negative splits. The Bonferroni p-value is computed by multiplying the raw p-value by the number of variables examined at the node and is used to account for the multiple testing. Eleven significant splits were found, and a high percentage, 79.5% (70/88), of the most active molecules were isolated in only three terminal nodes (shaded in gray with a bold border). For this data set, the analysis was performed in under 5 CPU seconds on a Silicon Graphics R10000 processor. The very small p-values should be noted. These confirm the information content of the descriptors and the value of large sample sizes.

When the RP classification tree is examined, it is often informative to see the distribution of potencies at a node and to see how a split at a node divides the distribution of potencies into the two daughter nodes. A nonparametric density plot is available to display the potency distribution at the node, with the potency distribution of the two daughter nodes. For example, the distributions of the parent node N01 and daughters N011 and N010 are shown in Figure 3. The density plot is computed by weighting each point by a Gaussian kernel function with a configurable bandwidth.[33] When the assay variability is known, the assay standard deviation can be directly used for the bandwidth.

To understand the chemical relevance of the splits obtained from recursive partitioning, a molecular structure viewer, ProjectView, was developed that not only displays structures, but highlights the portions of the molecules described in the rules. SCAM is a general statistical analysis program and is not limited to the display (or manipulation) of one type of descriptor, but rather passes individual descriptor rules to an external program which highlights the appropriate atoms or bonds. To SCAM, descriptors are just text strings, and it is up to external programs to interpret the results and display them accordingly. SCAM has an option that allows the user to reference a MDL SD-file containing the structures for the data set. Rather than reading each structure directly into memory, a list of "seek" indices is computed once for the SD-file since these files can be quite large. Then, whenever the user requests to see the compounds at a node, it is a simple matter of performing seeks to the appropriate offsets in the SD file to obtain the structures of interest. Further analysis of the structures in a given node can be performed since various structure manipulation options are available in ProjectView such as a substructure search algorithm.[34] For example, it is possible to select for viewing only those structures possessing a given substructure. The structures are then automatically rotated to best match the query substructure, thus facilitating visual structure/activity analysis. Sorting and selection of structures with a data value in a given range are also available. Representative structures found in nodes N1 and N011 are shown in Figure 4, with the structural features causing the classification highlighted. For clarity, these illustrations were redrawn.

Atom triples can also be used as descriptors. These are the topological equivalent of a geometric, three-point pharmacophores. (See Figure 1c for an example of an atom triple.) In the MAO data set, there are 125 175 unique triples and each structure is scored 0/1 for the absence/presence of these triples. The atom triple RP/SCAM decision tree is shown in Figure 5. The best feature that splits off 37 compounds is the presence of C(1,2)−8−C(2,1)−6−C(1,0)−5−. Note that C(2,1) is a terminal carbon with a triple bond. This feature is somewhat unique, and it occurs in pargyline. Node N1 is split twice, and 34 compounds with an average activity of 2.97 (out of 3.00) are found. Node N0 with 1613 compounds is split on the basis of the triple C(3,0)−2−N(2,0)−2−N(2,0)−3−. The N−N single bond is important. After one more split, 36 compounds with an average activity of 2.6 are found. Representative structures from these nodes are given in Figure 6. Note that the adjusted p-values for each split are small, 10^{-82} to 10^{-3}, indicating that these splits are unlikely to be chance findings and that the atom triples capture important information about the molecular features important for biological activity. Also note that the compounds in the active nodes are structurally distinct, which implies that there are different mechanisms for activity. The literature supports multiple mechanisms for MAO actitivity.[35] The SCAM algorithm and simple molecular descriptors have successfully found the two known mechanisms, binding to the active site and binding to a cofactor.

Virtual Screening. Once the analysis has been completed, a file describing the rules that create the RP classification tree can be written to disk, and a utility program, Pachinko,[36] can be invoked on a new data set to classify those structures. Thus, a screening set of compounds can be assayed, the results analyzed with SCAM, and an entire corporate collection, or a virtual library, literally "dropped down" the tree to suggest additional compounds for biological screening or synthesis. It is also possible to divide data into training and validation data sets and use Pachinko to test the

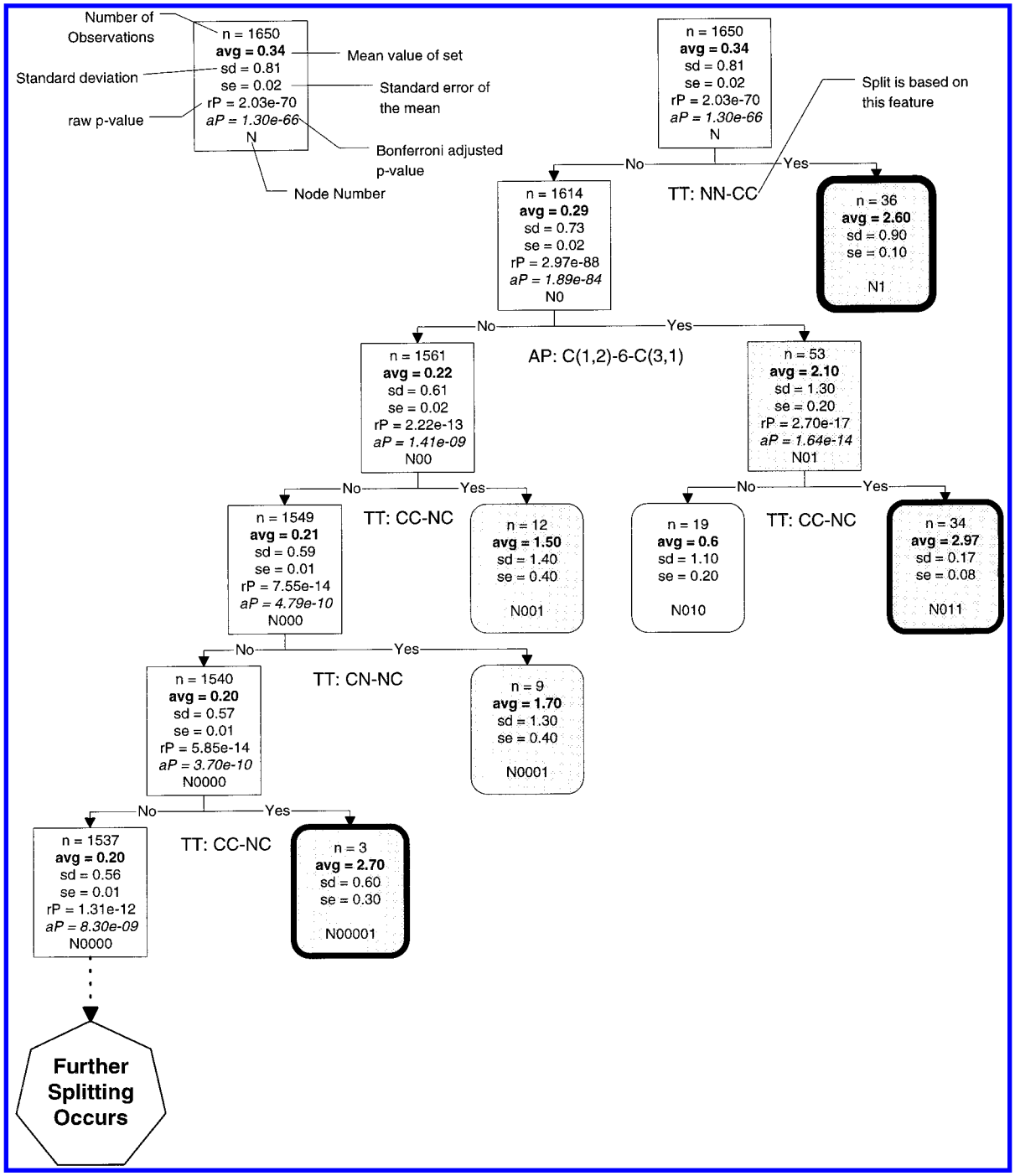


Figure 2. Atom pair/topological torsion RP/SCAM decision tree. Gray nodes indicate a collection of structures with high activity.

predictive powers of the tree.[22] Pachinko can predict the activity of approximately 100 structures every CPU second on a Silicon Graphics R10000 processor.

Quite often when using a large number of descriptor variables, there is more than one descriptor that would give exactly the same split at a node. These variables are perfectly correlated. When the variable associated with the most significant split has other perfectly correlated variables, all such descriptors are stored so that these rules can later be used for as input to the Pachinko program. In the training data set, all perfectly correlated variables will be found within the structures at a right node, though, in theory, possessing only one of these features might be necessary for biological activity. Within the Pachinko program, there is an option either to force all correlated variables to match for a rule to be satisfied, or else to have any one matching descriptor for the right path in a tree be taken. Pseudocode for Pachinko is given in Appendix B of the Supporting Information.

To test the predictive power of this analysis process, we use the RP tree from the atom pair, topological torsion analysis, and structures/activities from the World Drug Index (WDI). Several known MAO inhibitors that were in the training set were identified that were also in WDI, but several other known MAO inhibitors that were not in the training set were also found. Shown in Figure 7 are some of the structures predicted to be "active". There were 72 (0.2%) WDI structures classified as having MAO activity out of a possible 35 631 structures in the database. Using the rules from the RP tree, we predicted 227 structures would likely have MAO activity, and 7 of the 227 were classified as such

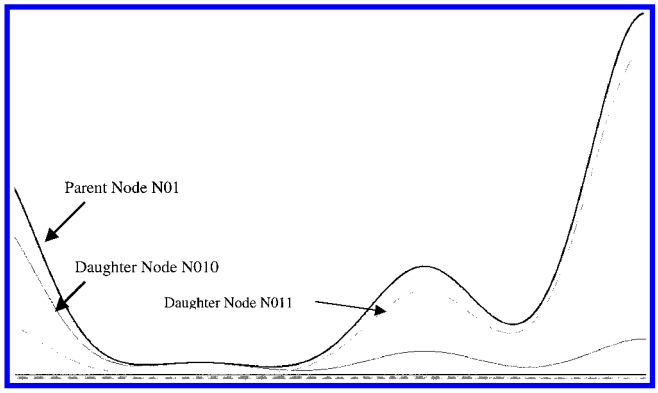


Figure 3. Distribution of potencies for parent (N01) and daughter (N011 and N010) nodes from the atom pair/topological torsion RP/SCAM decision tree. Daughter node N011 shows a definite increase in the number of active molecules versus daughter node N010.

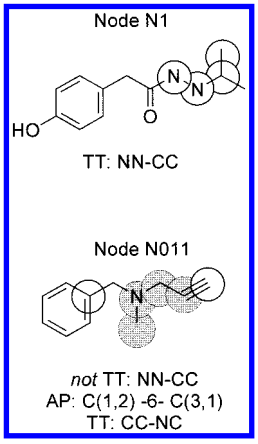


Figure 4. Representative structures found in nodes N1 and N011 of the atom pair/topological torsion RP/SCAM decision tree.

(3%). Thus, we see a 15-fold hit rate over random for this analysis (3%/0.2% = 15.4, p-value < 0.0001). Other CNS activities such as hypotensives or psychosedatives were reported for the 227 predicted activities, leading us to speculate that some of the structures might also have unreported MAO activity.

DISCUSSION

We are at an important juncture in the evolution of the drug discovery process. For the first time, data sets large enough to statistically determine the structural features relevant to biological activity are being routinely produced. Large chemical libraries, both general and focused, and large, heterogeneous sets of compounds are being tested. These data are being captured and stored and are readily available for analysis. Can this expensive data be turned into information to improve the efficiency of the drug discovery process?

The structure/activity analysis of large, structurally heterogeneous data sets is problematic for several reasons. First, descriptors must be computed for many structures, and the computational methods must be robust enough not to fail very often, if at all. The temporal value of a marketable drug can be one to several million dollars per day. Thus, the expedient delivery of new compounds to market is crucial to the success of pharmaceutical companies. With that in mind, it is not surprising that rapid, new-lead discovery techniques are now receiving great attention. This, of course, means that all data generation, collection, and analysis should be done as quickly as possible; in addition to simply being feasible, the computational process for statistical analysis of high-throughput screening data should be fast.

Any large-scale SAR statistical technique must deal with the distinct possibility that the active compounds might act through very different mechanisms. Ligands can bind in different modes and might even interact with different pockets to produce their effects, so although two compounds might have virtually identical binding values, these particular results may be the result of multiple, different binding modes or pockets and, thus, could depend on very different structural characteristics. Many standard statistical methods, multiple linear regression, partial least squares, etc., perform poorly or erratically when presented with such mixtures. Nonstandard methods such as neural networks are also expected to perform poorly under these circumstances. Consider an attempt to use simple linear regression to model a complex mixture. Suppose that there are two underlying mechanisms and that for one there is a linear increase in potency with increasing values of a particular descriptor. Further suppose that for the second mechanism potency decreases linearly with the same variable. Obviously, linear regression modeling can fail in this situation. This problem generalizes to many statistical methods employed for QSAR.

Previous attempts to determine quantitative structure/activity relationships have been usually limited to a congeneric series of compounds. Even here, there are often problems that require exclusion of structures that do not follow the mathematical model, lack the appropriate QSAR parameters, or require special mathematical tricks, e.g., introduction of indicator variables to note special circumstances. There is usually an admonition that the analysis method is expected to work as long as the compounds act through a single mechanism. Obviously, it is well recognized that most statistical methods require a single underlying process.

With high-throughput screening data, the assumption of a single mechanism of action for all the active compounds clearly cannot be relied upon. A simplistic solution to this problem is to select the most active compounds, ignoring the information contained in the remainder of the screening data. This is neither an efficient nor an effective use of the investment made in collecting the data. Cramer et al.[14a] were the first to attempt a formal statistical analysis of a large, heterogeneous structure data set. More recently, workers at Lederle[14b] used atom pairs and topological torsions to build a trend vector to predict the activity of untested compounds. The methods proceed as follows. First, use a training set where activities are known and compute the relative frequency that a fragment occurs in an active compound. Next, normalize the relative frequencies over all the fragments, under consideration. New compounds can be scored for the frequency of fragments and those that score well are predicted to be more likely to be active. However, there is no formal recognition in these methods of multiple mechanisms, and there is only the empirical observation that the methods are more successful than chance selection of compounds. The methods could be "fooled" in the following situation. Again suppose there are two underlying mechanisms of action for a hypothetical biological activity. If a particular fragment associated with activity is found in half of the active compounds in a data set and another fragment

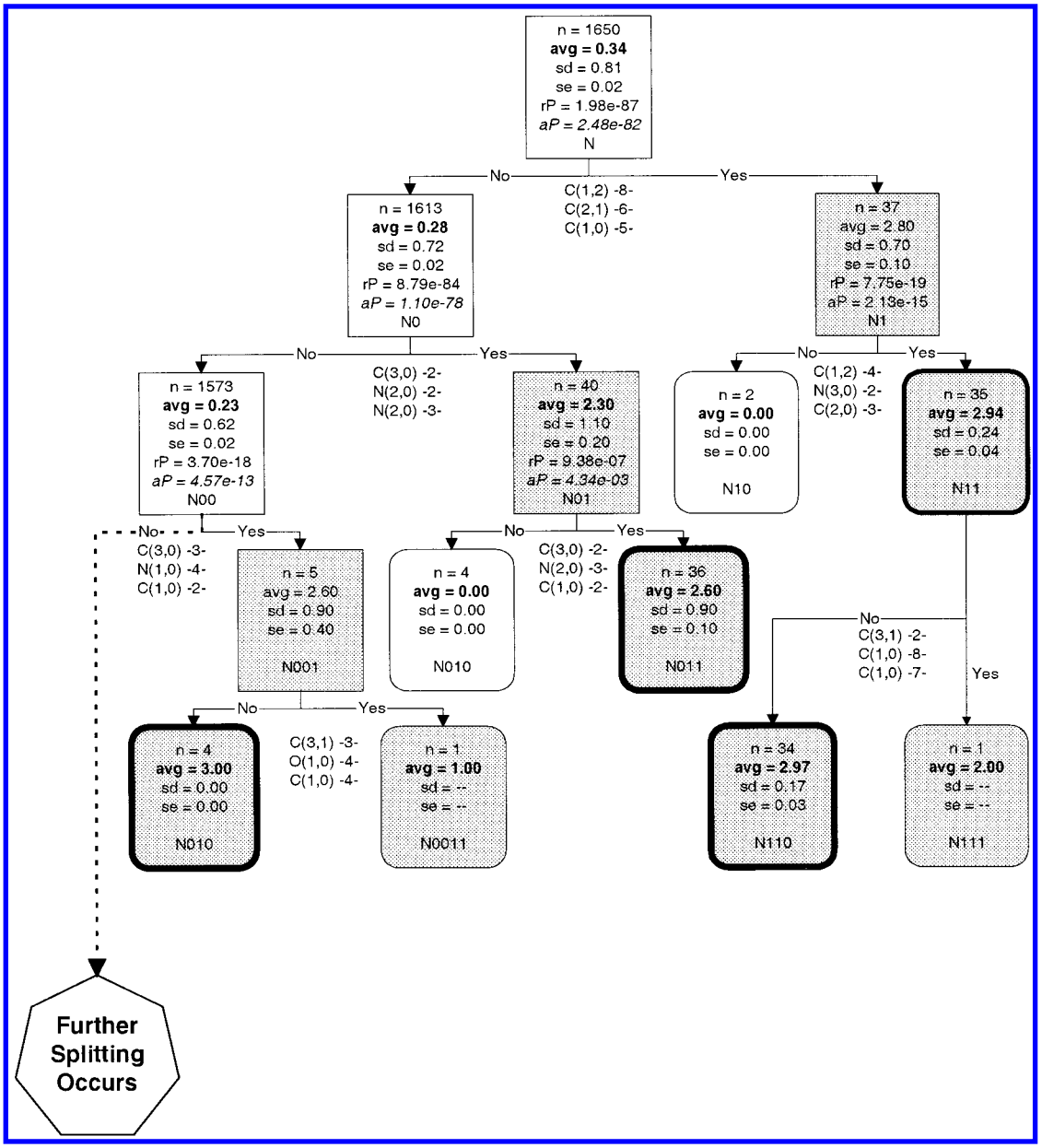


Figure 5. Atom triple RP/SCAM decision tree. Gray nodes indicate a collection of structures with high activity.

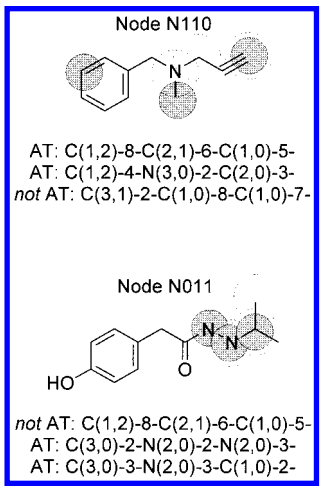


Figure 6. Representative structures found in nodes N110 and N011 of the atom triple RP/SCAM decision tree.

is found in the other half of the active molecules, then the relative frequency of each fragment among the active

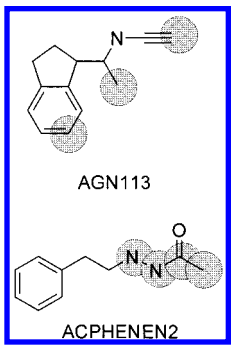


Figure 7. Example of MAO inhibitors in the World Drug Index predicted active by RP/SCAM.

structures is 0.50. In reality, the value is 1.00 within each active class; the other active class dilutes each relative frequency value! If there are many active classes, each requiring special fragments, then the critical relative frequency values for each class will be diluted by members of the other classes. If there are enough classes, then it may

appear that no single fragment is significant. Complicating this analysis is the fact that there are often bioisosteric fragments, i.e., different fragments that biochemically act the same, that somehow must be included in the analysis.

In contrast, recursive partitioning algorithms can deal with multiple mechanisms and, depending upon the types of features computed, bioisosteric replacements as well. When the algorithm makes a split of the data set on a feature, structures with the feature go down one arm of the tree, and if that feature is specific to a mechanism, then the compounds following that mechanism are effectively split from the bulk of the data. This phenomenon can clearly be demonstrated by examining the RP tree in Figure 2. The first split uses a topological torsion feature, NN−CC, to split 36 compounds with a mean activity of 2.60 from the rest of the data set. These compounds are irreversible, suicide inhibitors of MAO.[35] The N−N−C(=O) hydrolyzes to hydrazine and covalently binds to the protein. Alternatively, the 34 compounds in the N011 node have the typical features of a pargyline-like compound: a triple bond, a tertiary nitrogen, and an aromatic ring. These compounds are themselves suicide inhibitors and covalently attach to the MAO flavin cofactor.[35]

The CART recursive partitioning algorithm was applied to a structure/activity data set by Giampaolo et al.[37] This data set was relatively small, 56 compounds and 67 features, and the compounds were all nucleophiles and, arguably, expected to follow the same mechanism. The types of variables used in their analysis merit comment. The list is predominantly made up of "whole molecule" properties, e.g., molecular weight, solvated surface area, dipole moment, etc., so very specific, interacting features are largely not considered. Also, many of the features are experimentally determined, e.g., melting point, aqueous solubility, density, and these properties require that the compound be synthesized and measurements taken. Experimental measurements take time and involve some expense, so they are not feasible for large data sets, nor is this strategy amenable to a virtual screening paradigm. Also, whole molecule descriptors are more problematic for directing synthesis. The CART algorithm itself is computer intensive in that a large, bushy tree is built and then pruned back using cross validation. Analysis of a data set that might have hundreds of thousands of variables is not feasible using CART.

Machine learning methods have recently been applied to QSAR problems.[12] An inductive logic program, GOLEM, was used to analyze 44 trimethoprim analogues. A large number of features, on the order of thousands, can be examined, and the resulting SAR rules are clear in that they note specific features to maintain or to avoid. It is also clear that multiple mechanisms will give rise to multiple rules. This method is computer intensive and feasible for hundreds of compounds and thousands of features. Indeed, it is specifically noted that analysis of a data set with over a million features is not computationally feasible.

More recently, GOLEM[12] was used to predict rodent carcinogenicity. Here the compounds are structurally very diverse, and many different molecular mechanisms are operating. Even the biological end points change from study to study! A total of 330 compounds were used in the training set, and 39 new compounds were predicted. The predictions of GOLEM were the best among seven methods where only chemical structural information was used.

One fragment-based approach to automatically uncover SARs is the CASE/MULTICASE methodology.[38] This method uses all possible linear fragments of 2−10 connected heavy atoms as descriptors and so is similar to the use of topological torsions. Analysis proceeds by finding the fragment that best separates out a group of active compounds. Those compounds form a class and are removed. This step is repeated on the null group. Then standard variables and linear regression are used to develop a QSAR for each identified class, compounds that contain a "biophore."

Several questions can be raised about the CASE/MULTICASE methodology. First, the response variable must be binomial which entails dividing the compounds into those considered active and inactive. Where to make this "cut" in a data set is problematic; inevitably some information is lost using a binomial variable rather than a continuous measurement of activity. Additionally, there are a considerable number of fragments of size 2−10, so there needs to be some adjustment for multiple testing. There is some adjustment for chance correlations in the CASE/MULTICASE methodology, but better statistical methods are available.[39]

More subtle problems exist as well. Several fragments might be biologically equivalent, possibly leading to an unnecessary subdivision of the compounds. For example, −OH and $-NH_2$ are often considered to be bioisosteres, so fragments that contain these two groups might better form one class rather than two. Furthermore, more than one noncontiguous fragment might be required for activity, and this technique is designed to primarily recognize one fragment per class. The CASE algorithm fared poorly when compared to other prediction methods.[40]

A recent report[41] on the use of MCASE to analyze an NCI HIV data set of 24 110 compounds contains the following statement: "When the total number of chemicals of the learning set is large, millions of fragments may be generated. The memory required and CPU time needed to manage such a vast array of data are beyond the current capability of most computers." They then engaged in a complex compound selection process to reduce the number of compounds examined to 1819 so they could apply the MCASE algorithm.

RP/SCAM is successful for several reasons. First, much of the interaction of a compound with a biological system seems to be captured by a relatively few, sharply defined features. The difficult question is how to find these few features from a great many. Second, SCAM *can* separate out classes of compounds that are acting through different mechanisms although there is no guarantee that it will. Alternative statistical methods, multiple linear regression, partial least squares, neural networks, etc., often fail when confronted with mixtures of models. Third, any single structure does not have many of the vast number descriptors that are possible for a compound. The descriptor matrix is very sparse, so sparse matrix techniques are used to optimize performance and computer memory utilization. Furthermore, the algorithms are sufficiently fast to explore large search spaces interactively. Finally, the tree display of the results is very clear and evocative. Expert medicinal chemists follow the analysis results quickly and usually suggest new hypotheses for testing.

Although the advantages of RP in general, and SCAM in particular, are obvious, some limitations should be noted.

First, RP is data greedy, requiring large data sets, 300 or more observations depending on the complexity of the process. Sample size decreases drastically as splits are made, and the statistical power dwindles concomitantly. This is not a problem for most companies now with the universal adoption of high-throughout screening. Second, higher order interactions can be masked by nonsignificant main effects where two or more features perfectly counterbalance one another, so an initial split is not made. Third, disposition of outliers (or in this case, single molecules) is difficult; the proper categorization of structures with features that occur once or only a few times in the data set is problematic with this or any statistical technique. Fourth, occasionally the user can be fooled by the tree's simplistic form and might miss highly correlated, but more relevant features. As is the case with any mechanical algorithm, experts should carefully examine the results. Furthermore, there are no formal mechanisms for the statistical design of followup experiments, although several methods are under consideration in our laboratory. Finally, binary, yes/no, feature variables may be too crude in some instances to allow a meaningful analysis.

CONCLUSIONS

RP/SCAM has several clear advantages over current QSAR methods for large, heterogeneous structure/activity data sets. The code is very fast even when dealing with seemingly impossibly sized data sets. Internally, we have analyzed data sets with over 100 000 compounds and 2 000 000 descriptors in less than an hour. In addition, SCAM can deal effectively with mixtures since the determination of a QSAR for a mixture of mechanisms undoubtedly requires many examples to find important features. Proper statistical corrections are needed to limit false associations. Furthermore, the ability to handle many descriptors eliminates the need for experts to pare down descriptor lists. Experts can focus their attention on *expanding* lists of plausible descriptors, increasing the likelihood of developing a chemically satisfying explanation of the data. An obvious extension of this work is to employ 3D geometric features as potential rules[42,43] wherein RP/SCAM can be used to generate 3D pharmacophores. Finally, RP/SCAM is a general data mining method and can be applied to other areas of interest where large data sets are accumulating.

DATA

The MAO data set is available from Abbott Laboratories: contact Daniel W. Norbeck, Abbott Laboratories, 100 Abbott Park Rd., Abbott Park, IL 60064-3500. The World Drug Index is available from Daylight Chemical Information Systems, Santa Fe, NM.

PATENTS

There are patents pending covering certain aspects of this research.

Supporting Information Available: Pseudocode to generate an RP/SCAM decision tree (Appendix A) and pseudocode for automatic structure prediction using Pachinko (Appendix B). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Presented in part at the 213th National American Chemical Society Meeting, San Francisco, CA, 1997.

(2) (a) Gallop, M. A.; Barrett, R. W.; Dower, W. J.; Bodor, S. P. A.; Gordon, E. M. Applications of Combinatorial Technologies to Drug Discovery. 1. Background and Peptide Combinatorial Libraries. *J. Med. Chem.* **1994**, *37*, 1233−1251. (b) Gordon, E. M.; Barrett, R. W.; Dower, W. J.; Fodor, S. P. A.; Gallop, M. A. Applications of Combinatorial Technologies to Drug Discovery. 2. Combinatorial Organic Synthesis, Library Screening Strategies, and Future Directions. *J. Med. Chem.* **1994**, *37*, 1385−1401. (c) Williard, X.; Pop, I.; Bourel, L.; Horvath, D.; Baudelle, R.; Melnyk, P.; Deprez, B.; Tartar, A. Combinatorial chemistry: a rational approach to chemical diversity. *Eur. J. Med. Chem.* **1996**, *31*, 87−98.

(3) Presented in part at the 1996 MDL User Conference, May 6, 1996, Philadelphia, PA.

(4) (a) OMG, www.oxmol.com. (b) MDL, www.mdli.com. (c) Tripos, www.tripos.com. (d) MSI, www.msi.com.

(5) Kuntz, I. D.; Meng, E. C.; Shoichet, B. K. Structure-Based Molecular Design. *Acc. Chem. Res.* **1994**, *27*, 117−123.

(6) Martin, Y. C.; Bures, M. G.; Danaher, E. A.; DeLazzer, J.; Lico, I.; Pavlik, P. A. A Fast New Approach to Pharmacophore Mapping and Its Application to Dopaminergiic and Benzodiazepine Agonists. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 83−102.

(7) (a) Raman, V. K. Applications of Artificial Intelligence in Chemistry. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 937 (b) Bolis, G.; Di Pace, L.; Fabrocini, F. A machine learning approach to computer-aided molecular design. *J. Comput.-Aided Mol. Des.* **1991**, *5*, 617−628. (c) Cohen, A. A.; Shatzmiller, S. E. Implementation of artificial intelligence for automatic drug design. I. Stepwise computation of the interactive drug-design sequence. *J. Comput. Chem.* **1994**, *15*, 1393−1402.

(8) (a) Corey, E. J. Computer-assisted analysis of complex synthetic problems. *Q. Rev., Chem. Soc.* **1971**, *25*, 455−482. (b) Salatin, T. D.; Jorgensen, W. L. Computer-assisted mechanistic evaluation of organic reactions. 1. Overview. *J. Org. Chem.* **1980**, *45*, 2043−2051.

(9) (a) Peris, M. An overview of recent expert system applications in analytical chemistry. *Crit. Rev. Anal. Chem.* **1996**, *26*, 219−237. (b) Carhart, R. E.; Smith, D. H.; Brown, H.; Djerassi, C. Applications of artificial intelligence for chemical inference. XVII. Approach to computer-assisted elucidation of molecular structure. *J. Am. Chem. Soc.* **1975**, *97*, 5755−5762. (c) Pomeroy, R. S.; Kolczynski, J. D.; Denton, M. B. Information-based expert systems for atomic emission spectroscopy. *Appl. Spectrosc.* **1991**, *45*, 1111−1119. (d) Luinge, H. J.; Kleywegt, G J.; Van't Klooster, H. A.; Van der Maas, J. H. Artificial intelligence used for the interpretation of combined spectral data. 3. Automated generation of interpretation rules for infrared spectral data. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 95−99.

(10) (a) Gasteiger, J.; Zupan, J. Neural networks in chemistry. *Angew. Chem., Int. Ed. Engl.* **1993**, *32*, 503−536. (b) Burden, F. R. Using artificial neural networks to predict biological activity from simple molecular structural considerations. *Quant. Struct.-Act. Relat.* **1996**, *15*, 7−11. (c) 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.

(11) Moriguchi, I.; Hirono, S.; Liu, Q.; Nakagome, I. Fuzzy adaptive least squares and its application to structure−activity studies. *Quant. Struct.-Act. Relat.* **1992**, *11*, 325−331.

(12) (a) King, R. D.; Srinivasan, A. The discovery of indicator variables for QSAR using inductive logic programming. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 571−580. (b) King, R. D.; Muggleton, S.; Lewis, R. A.; Sternberg, M. J. Drug Design by Machine Learning: the Use of Inductive Logic Programming to Model the Structure-Activity Relationships of Trimethoprim Analogues Binding to Dihydrofolate Reductase. *PNAS* **1992**, *89*, 11322−11326.

(13) Bolis, G.; Di Pace, L.; Fabrocini, F. A machine learning approach to computer-aided molecular design. *J. Comput.-Aided Mol. Des.* **1991**, *5*, 617−628.

(14) (a) Cramer, R. D.; Redl, G.; Berkoff, C. E. Substructural analysis. Novel approach to the problem of drug design. *J. Med. Chem.* **1974**, *17*, 533−535. (b) Hodes, L.; Hazard, G. F.; Geran, R. I.; Richman, S. A statistical-heuristic method for automated selection of drugs for screening. *J. Med. Chem.* **1977**, *20*, 469-475. (c) 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.

(15) Fayyad, U.; Piatetsky-Shapiro, G. Smyth, P. *From Data Mining to Knowledge Discovery*; AAI Press/MIT Press: Cambridge, MA, 1995.

(16) http://www.kdnuggets.com.

(17) Weber, L.; Wallbaum, S.; Broger, C.; Gubernator, K. Optimization of the biological activity of combinatorial compound libraries by a genetic algorithm. *Angew. Chem., Int. Ed. Engl.* **1995**, *34*, 2280−2282.

(18) Singh, J.; Ator, M. A.; Jaeger, E. P.; Allen, M. P.; Whipple, D. A.; Soloweij, J. E.; Chowdhary, S.; Treasurywala, A. M. Application of Genetic Algorithms to Combinatorial Synthesis: A Computational

Approach to Lead Identification and Lead Optimization. *J. Am. Chem. Soc.* **1996**, *118*, 1669−1676.
(19) Gobbi, A.; Poppinger, D.; Rohde, B. First International Electronic Conference on Synthetic Organic Chemistry, 1997, http://www.unibas.ch/mdpi/ecsoc/f0008/f0008.htm.
(20) Myers, P. L.; Greene, J. W.; Saunders: J.; Teig, S. L. Rapid, reliable drug discovery. *Today's Chem. Work* **1997**, *6*, 45−53.
(21) Young S. S.; Hawkins D. M. Analysis of a 2^9 Full Factorial Chemical Library. *J. Med. Chem.* **1995**, *38*, 2784−2788.
(22) Hawkins, D. M.; Young, S. S.; Rusinko, A., III. Analysis of a Large Structure−Activity Data Set Using Recursive Partitioning. *Quant. Struct.-Act. Relat.* **1997**, *16*, 296−302.
(23) (a) 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. (b) Hawkins, D. M. *FIRM Formal Inference-based Recursive Modeling*, Release 2; University of Minnesota: St. Paul, MN, 1995.
(24) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees*; Wadsworth: New York, 1984.
(25) Quinlan, J. R. *C4.5 Programs for Machine Learning*, Morgan Kaufmann Publishers: San Mate, CA, 1992.
(26) Brown, R. D.; Martin, Y. C. Use of Structure−Activity Data to Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.
(27) Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. Topological Torsion: a New Molecular Descriptor for SAR Applications. Comparison with Other Descriptors. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 82−85.
(28) Fisanick, W.; Coss, K. P.; Forman, J. C.; Rusinko, A., III. Experimental system for similarity and 3D searching of CAS registry substances. 1. 3D substructure searching. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 548−559.
(29) Good, A. C.; Kuntz, I. D. Investigating the extension of pairwise distance pharmacophore measures to triplet-based descriptors. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 373−379.
(30) Good, A. C.; Ewing, T. J. A.; Gschwend, D. A.; Kuntz, I. D. New molecular shape descriptors: application in database screening. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 1−12.
(31) Morgan, J. A.; Sonquest, J. N. Problems in the analysis of survey data, and a proposal. *J. Am. Stat. Assoc.* **1963**, *58*, 415−434.
(32) (a) Wouters, J., Structural aspects of monoamine oxidase and its reversible inhibition. *Curr. Med. Chem.* **1998**, *5*, 137−162. (b) Gareri, P.; Stilo, G.; Bevacqua, I.; Mattace, R.; Ferreri, G.; De Sarro, G. Antidepressant drugs in the elderly. *Gen. Pharmacol.* **1998**, *30*, 465−475.
(33) (a) Silverman, B. W. *Density Estimation in Statistics and Data Analysis*; Chapman & Hall: London, 1986. (b) Terrell, G. R. *J. Am. Stat. Assoc.* **1990**, *85*, 470.
(34) Ullman, J. R. An algorithm for subgraph isomorphism. *J. Assoc. Comput. Mach.* **1976**, *23*, 31−42.
(35) (a) Nelson, S. D.; Mitchell, J. R.; Timbrell, J. A.; Snodgrass, W. R.; Corcoran, G. B., III. Soniazid and iproniazid: activation of metabolites to toxic intermediates in man and rat. *Science* **1976**, *193*, 901−903. (b) Maycock, A. L.; Abeles, R. H.; Salach, J. I.; Singer, T. P. The structure of the covalent adduct formed by the interaction of 3-(dimethylamino)-1-propyne and the flavine of mitochondrial amine oxidase. *Biochemistry* **1976**, *15*, 114−125.
(36) Pachinko is a Japanese parlor game in which a steel ball bounces down through a series of pins ultimately falling into a bin at the bottom of the board. Each yes/no decision in the RP tree corresponds to a pin in the Pachinko game. Whereas Pachinko is largely random, the RP tree is deterministic on the basis of the rules generated from the analysis algorithm.
(37) Giampaolo, C.; Gray, A. T.; Olshen, R. A.; Szabo, S. Predicting chemically induced duodenal ulcer and adrenal necrosis with classification trees. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 6298−6302.
(38) (a) Klopman, G. Artificial intelligence approach to structure−activity studies. Computer automated structure evaluation of biological activity of organic molecules. *J. Am. Chem. Soc.* **1985**, *106*, 7315−7321. (b) Klopman, G. MULTICASE. 1. A Hierarchical Computer Automated Structure Evaluation Program. *Quant. Struct.-Act. Relat.* **1992**, *11*, 176−184. (c) Klopman, G.; Li, J.-Y. Quantitative structure-agonist activity relationship of capsaicin analogues. *J. Comput.-Aided Mol. Design* **1995**, *9*, 283−294.
(39) Miller, R. G. *Simulatneous Statistical Inference*; Springer-Verlag: New York, 1981.
(40) Downs, G. M.; Gill, G. S.; Willett, P.; Walsh, P. Automated descriptor selection and hyperstructure generation to assist SAR studies. *QSAR Environ. Res.* **1995**, 253−264.
(41) Klopman, G.; Tu, M. Diversity analysis of 14 156 molecules tested by the National Cancer Institute for anti-HIV activity using the quantitative structure−activity relationship expert system MCASE. *J. Med. Chem.* **1999**, *42*, 992−998.
(42) Chen, X.; Rusinko, A.; Young, S. S. Recursive partitioning analysis of a large structure−activity data set using three-dimensional descriptors. *J. Chem. Inf. Comp Sci.* **1998**, *38*, 1054−1062.
(43) Chen, X.; Rusinko, A., III; Tropsha, A.; Young, S. S. Automated pharmacophore identification for large chemical data sets. *J. Chem. Inf. Comp Sci.* **1999**, *39*, 887−896.

CI9903049