NISS

A Sequential Approach for Identifying Lead Compounds in Large Chemical Databases

Markus Abt, Yong Bin Lim, Jerome Sacks,
Minge Xie, and S. Stanley Young

Technical Report Number 105
March, 2000

National Institute of Statistical Sciences
19 T. W. Alexander Drive
PO Box 14006
Research Triangle Park, NC 27709-4006
www.niss.org

A Sequential Approach for Identifying Lead Compounds in Large Chemical Databases

Markus Abt, YongBin Lim, Jerome Sacks, Minge Xie and S. Stanley Young[1]

Abstract. At the early stage of drug discovery, many thousands of chemical compounds can be synthesized and tested (assayed) for potency (activity) with High Throughput Screening (HTS). With ever increasing numbers of compounds to be tested (now often in the neighborhood of 500,000) it remains a challenge to find strategies via sequential design that reduce costs while locating classes of active compounds.

Initial screening of a modest number of selected compounds (first-stage) is used to construct a structure-activity relationship (SAR). Based on this model, a second stage sample is selected, the SAR updated and, if no more sampling is done, the activities of not yet tested compounds are predicted. Instead of stopping, the SAR could be used to determine another stage of sampling after which the SAR is updated and the process repeated.

We use existing data on the potency and chemical structure of 70223 compounds to investigate various sequential testing schemes. Evidence on two assays supports the conclusion that a rather small number of samples selected according to the proposed scheme can more than triple the rate at which active compounds are identified, and also produce SARs effective for identifying chemical structure. A different set of 52883 compounds is used to confirm our findings.

One surprising conclusion of the study is that the selection of the initial sample stage may be unimportant: Random selection or systematic methods based on chemical structures are equally effective.

Key words and phases: Combinatorial chemistry, data mining, high throughput screening, recursive partitioning, sequential design, structure-activity relationship.

1. INTRODUCTION

The search for a new drug to combat a disease begins with the development of an understanding about how the disease manifests itself on a molecular level. Once the molecular target, typically in form of a protein, has been identified, biological assays are developed that allow the testing (screening) of compounds with respect to their ability to interact with the protein. For this purpose automated screening systems are

---

[1] *M. Abt was, during the time of this research, Junior Research Fellow at the National Institute of Statistical Sciences and is now working in the department of Biostatistics at F. Hoffmann-La Roche AG, 4070 Basel, Switzerland. Y.B. Lim is Professor, Department of Statistics, Ewha Womans University, Seoul 120-750, Korea. J. Sacks is Senior Fellow, National Institute of Statistical Sciences, P.O. Box 14006, Research Triangle Park, NC 27709-4006. M. Xie is Assistant Professor, Department of Statistics, Rutgers University, Piscataway, NJ 08855. S. S. Young is Principal Consultant, GlaxoWellcome Inc., Five Moore Drive, Research Triangle Park, NC 27709. This research was supported in part by NSF-DMS 9208758 and 9700867. Lim's research was also supported in part by the Korea Science & Engineering Foundation Grant 981-0105-024-2.*

available that, depending on the assay, allow the screening of hundreds to thousands of compounds a day. The search is for "lead" compounds which eventually can be modified to produce new and effective drugs.

Corporate chemical databases of compounds that are available for testing can contain hundreds of thousands of molecules. In addition, compounds are available from commercial sources or can be obtained through combinatorial synthesis from elementary building blocks (Cortese, 1996). Even larger can be the number of compounds in virtual libraries, a collection of theoretically possible but not yet synthesized molecules. The introduction of High Throughput Screening (HTS) allows the testing of large numbers of compounds in comparatively short time. Exhaustive screening of compound collections, despite miniaturization efforts (Burbaum, 1998), is impractical in view of the ever increasing size of the collections. A systematic approach via a sequential search that tests a comparatively small number of molecules in an inventory and identifies structural features that might then guide the selection process towards selecting more effective compounds is therefore of great practical value. In order to explore such strategies, we used historic data from the complete screening of two different compound libraries in two different assays. See Figure 1.

Such sequential search schemes in the context of drug discovery face a number of daunting obstacles. First, the size of the space of compounds to be searched is in the tens of thousands at a minimum and can be in the millions to billions for virtual libraries. Second, the spaces themselves are highly complex. A molecule may be described at many levels of "accuracy", ranging from comparatively simple topological descriptions of dimension in the thousands to difficult to compute but fewer descriptions arising from quantum chemistry calculations. Third, the number of compounds in the space that have adequately high potency is very small, typically less than 0.5%. Fourth, the target of the search is not completely precise: The search is not only for high potencies but also for a variety of chemical structures associated with high potency that medicinal chemists can use to take follow-up steps in synthesizing new molecules. The chemists need multiple chemical classes as tractable starting points for synthetic modification because compounds, besides being potent,

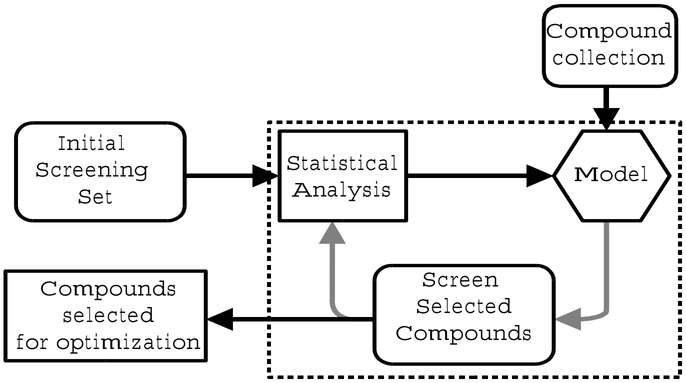

Fig. 1. *Sequential screening process. An initial compound set is screened and statistically analyzed giving a model that describes compound features associated with activity. These structure-activity rules are used to select additional molecules for screening from the available compound collection. The combined data set is again statistically analyzed. The cycle is repeated until compounds are selected for atom-by-atom optimization.*

need to meet requirements about toxicity, side effects, duration of effect and specificity. Roughly, what has to be faced is a problem of searching a potency surface over a large discrete space of very high dimension for a variety of high peaks.

Further constraints arise because of practical considerations. High setup costs at each stage preclude a purely sequential scheme so we are limited to a few number of stages. Being first to market a new drug can lead to gains in the millions of dollars per day; this value of time imposes limitations on computational and data analytic strategies.

Several fundamental statistical issues are to be faced when implementing a sequential scheme. In order to start the search, compounds must be selected for testing in a first stage. At subsequent stages, the selection of additional compounds is based on the potencies found in testing the compounds selected at previous stages. Implicit is the use of the data to develop a relationship between the geometry of the space (the structure of the molecules) and the biological activity (the potency measured by the assay). The chemistry and the geometry are intertwined and affect the development of useful structure-activity relationships (SAR). Further questions arise about the number of stages that are needed as well as the number of compounds to be selected at each stage.

We describe a sequential scheme that takes up these basic issues via a specific case study involving a space of 70223 compounds with known activities. The goal is to propose a strategy that is both effective and can be implemented in practice. The issues mentioned above are discussed and answers proposed and confirmed in additional examples and studies.

Our conclusions, stated succinctly, are:

- Use a sequential approach;
- Design of a first stage sample is unimportant — random selection is hard to beat;
- Careful design of next stages is advantageous;
- Two stages are enough.

We have not attempted to explore the challenging question of whether a sequential design problem can be precisely formulated and analyzed in our setting. The barriers to doing so are formidable, not the least being the large dimension of the space of descriptors and the interactions among them.

Section 2 describes the initial data set and chemical features for the case study. All the potencies are available for this data set. But we proceed in ignorance of this fact until the final step of our study where we use the unselected compounds to validate the procedures and compute their performance characteristics. In Section 3 we review recursive partitioning and the particular statistical classification method used to establish the structure-activity relationship. Of prime importance is that the methods run very rapidly on large numbers of compounds, each described by a large dimensional vector of descriptors. The factors expected to be influential on the performance of our sequential scheme are discussed in Section 4. The main questions to be explored are how and how many compounds to select at the various stages of the sequential search scheme. Section 5 provides the layout of our initial experiment to identify a good screening strategy and Section 6 gives the analysis of the results. Some additional analyses exploring specific questions are discussed in Section 7. Confirmation experiments are described in Section 8. Section 9 discusses some future directions of investigation and also provides references to alternative approaches for modelling structure-activity relationships. Concluding remarks follow in Section 10.

As a result of our investigations, sequential screening is now routinely used by GlaxoWellcome scientists and has helped shorten the time needed to provide the medicinal chemists with interesting chemical structures for further optimization. Central to the approach is classification via recursive partitioning. The design (selection of compounds) at each of the individual stages appears to be of relatively minor importance. Results from applying the described approach to screening over 160,000 compounds were recently reported in Jones-Hertzog, Mukhopadhyay, Keefer, and Young (2000).

2. THE TESTBED DATA

The data set used as a testbed for the methods contains the potencies together with a description of the chemical structure for each of 70223 compounds from an assay carried out by GlaxoWellcome scientists.

The assay measures the potency of each of the 70223 compounds by recording their ability to bind to a protein and displace a specific (to the assay) standard compound that naturally binds in a cleft of the protein. The analogy of a “lock and key” is suggestive of the binding process. When a compound binds to a protein, a recordable color change can be observed. The intensity of this color change is a measure of the compound’s ability to bind to the protein; this intensity defines potency.

Generally, the uncertainty associated with measuring the potency of a compound will affect the performance of any discovery procedure. We will not attempt to take this into account below. Precision is to some extent sacrificed for speed and easy logistics.

It is common to apply a logarithmic transformation to the data in order to reduce the skewness of the distribution and to remove a possible dependence between the mean and the variability in the response. Some characteristics of the 70223 log potencies - we shall hereafter use log potency as the measure of potency - are listed in Table 1.

TABLE 1

Summary statistics of log potencies. The 25% and 75% quantiles are denoted by q_{25} and q_{75}, respectively.

min	q_{25}	median	q_{75}	max	mean	stdv
-1.190	0.837	0.918	1.012	3.102	0.927	0.166

The chemical structure of molecules can be described in several ways. A very basic description consists of the list of atoms that constitute the compound. Alternatively, we can use counts of fragments or functional groups referring to entities of atoms. For our study we used a topological descriptor based on atom pairs, see Carhart, Smith, and Venkataraghavan (1985): For any two non-hydrogen atoms Atom 1 and Atom 2 there typically are many paths of successive bonds and atoms in the compound that link Atom 1 and Atom 2. A path with the fewest number of atoms between Atom 1 and Atom 2 is called a minimal path and the number of atoms in such a minimal path is the topological distance between Atom 1 and Atom 2. Each atom pair is then of the form < Atom 1 description > - < topological distance > - < Atom 2 description >. The description of an atom consists of the elemental name, the number of bonds attached to it as well as the number of π-electrons in these bonds. For example, the description of a carbon atom which is attached to two non-hydrogen atoms and shares one π-electron with its neighbors is denoted by C(2,1). Thus, even atoms of the same type, two carbons, for example, are distinguished if they differ in the number of bonds attached to them and the number of π-electrons.

Although possible, multiple occurrences of the same atom pair in a molecule are not accounted for in our descriptors. Among the 70223 molecules, 8189 different atom pairs were found. The resulting molecular descriptions are then bitstrings of length 8189, where one and zero indicate the presence or absence of the corresponding atom pair. These vectors were produced using software developed by A. Rusinko based upon algorithms given in Carhart et al. (1985). Being able to rapidly produce descriptors that capture the important features of the chemical structure of a molecule is important. We cannot afford to take physical measurements to characterize compounds.

The number of atom pairs that occur in a compound varies greatly, see Figure 2. There are a few compounds with many atom pairs and a few with a small number of atom pairs. Atom pairs that occur in all compounds are not included; such pairs provide no information. The “biggest” compound contains 603 of the total 8189 atom pairs. For most compounds, the number of atom pairs is in the range from 80 to 150. There are atom pairs that occur in very few compounds; there are other atom pairs that appear in over 50000 in the set of 70223.

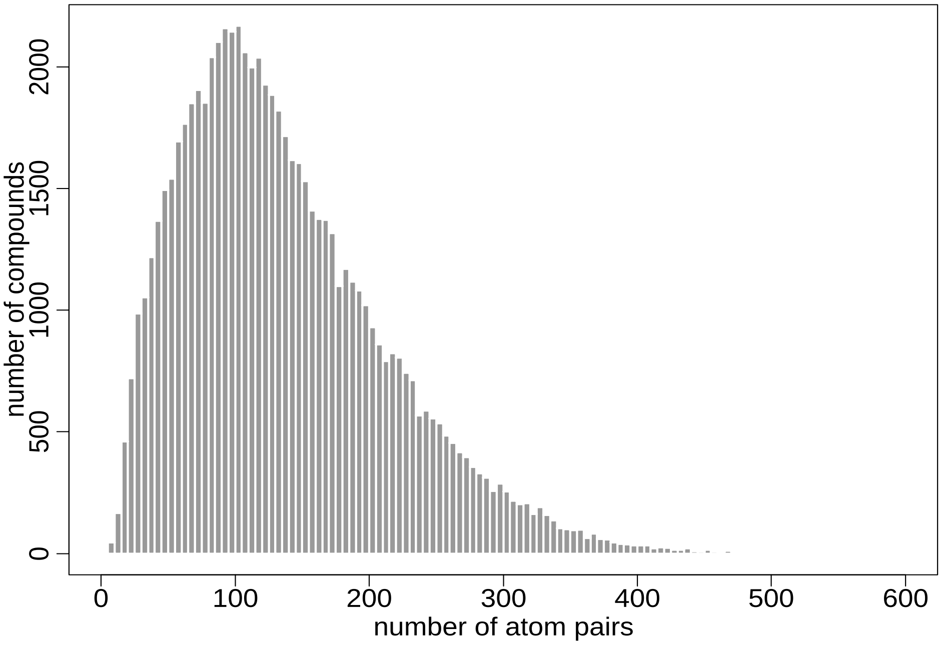

Fig. 2. *Number of compounds containing a given number of atom pairs.*

Another readily computed descriptor for a molecule is its Burden number, introduced by Burden (1989). The Burden number is a property of the connectivity matrix of a compound. The definition of this matrix starts by (arbitrarily) numbering the n non-hydrogen atoms occurring in the structure $1, \ldots, n$. Then, an $n \times n$-matrix is formed containing on its diagonal the atomic number of each atom, i.e., the number of protons in the atomic nucleus. The off diagonal elements are chosen as positive real numbers that depend on whether two atoms are neighbors and, if so, on the type of bond between them. Finally, the Burden number is defined as the smallest eigenvalue of this connectivity matrix. While other eigenvalues may also be useful, we will only consider the smallest one.

Though a relatively "coarse" description of a molecule, the Burden numbers are attractive because of their one-dimensional nature and the comparative ease of their computation. Moreover, two molecules with close Burden numbers often appear similar when comparing their chemical structures for example, by comparing numbers of fragments or functional groups two molecules have and have not in common.

While it is relatively cheap and easy to compute the descriptors of chemical structures of compounds, it is extremely expensive and time consuming to measure the potencies of an entire collection of chemical compounds. The above testbed data set is one of a few rare occasions where all compounds were tested. A very practical question is whether it is possible to find most of the potent chemicals by testing only a proportion of the compounds in a collection. The testbed data set is used to demonstrate one of such strategies in this paper. Thus, in our proposed sequential procedures (see Section 4), we proceed in ignorance of the potency value of a compound unless it is selected. Potency values of unselected compounds will be used only in the final step when we validate the procedures.

3. RECURSIVE PARTITIONING

The analysis of data sets with over 70000 observations and about 8000 independent variables is a formidable computational task. The underlying relationship between the response (potency) and the independent variables (atom pairs) could involve nonlinearities, thresholds, and interactions among the explanatory variables. Other complications result from the possibility that compounds may bind in different ways: some compounds in the data set may act through one mechanism while others act through a different mechanism. Classical methods such as regression analysis, stepwise regression, or principal components regression are likely to be compromised in these circumstances depending on how the predictor variables are chosen.

A less parametric method, capable of identifying important structural variables and their interactions, is recursive partitioning, a tree structured approach to regression and classification. The observations are partitioned by a sequence of splits (using the independent variables) resulting in a set of terminal nodes. The path leading to each terminal node reveals structural information about the compounds living in that node. This structural information can then be associated with the specific molecular features that divide the compounds into activity classes.

FIRM (Formal Inference-based Recursive Modeling) was proposed by Hawkins and Kass (1982) and is a recursive partitioning algorithm based on hypothesis testing (see also Hawkins, 1994). The algorithm is fast and can be modified to analyze large numbers of descriptors. In our case the explanatory variables are binary and the data matrix consisting of 70223 rows (corresponding to compounds) and 8189 columns (corresponding to atom pairs) is sparse. According to Figure 2, most compounds have fewer than 250 atom pairs and thus most rows will contain fewer than 250 ones. The sparsity of the matrix enabled Rusinko, Farmen, Lambert, Brown, and Young (1999) to develop specialized software, Statistical Classification of Activities of Molecules (SCAM), for rapid computation of a recursive partitioning.

Other versions of recursive partitioning have been implemented in the literature. Most notable of these is CART (Classification And Regression Trees) by Breiman, Friedman, Olshen, and Stone (1984), which can be applied to both continuous and categorical response data sets. CART relies on sophisticated cross-validation and pruning techniques to determine the size of the final tree and its terminal nodes. The very general nature of CART makes it a very flexible tool that can be used in a wide variety of applications, but might not be the most efficient choice for the types of data we are working with. In addition, SCAM has a built-in utility that allows the medicinal chemist to interactively view the chemcical structures of molecules.

Detailed comparisons of CART and SCAM with respect to computing time are still outstanding.

SCAM uses a simple t-test splitting criterion to select a binary split at every intermediate node. The t-test is done with a Bonferroni adjusted p-value (see Westfall and Young, 1993, for example) to protect against excessive splitting resulting from the multitude of possible splits. The resulting SCAM tree looks like a CART tree with binary splits. The criterion for the best split of a node is similar, but the pruning mechanisms and stopping rules are different.

As an example, consider 10000 compounds selected randomly from all 70223 molecules. The result of recursive partitioning applied to this example is displayed as a tree in Figure 3. The first step of the algorithm splits the data set into two groups according to the absence (left branch) or presence (right branch) of atom pair AP1. Based on a two sample t-test, the Bonferroni adjusted p-value associated with this split is 2.25E-6. Splits are called significant when the adjusted p-value is below 0.01. Splits on the same atom pair are possible in different parts of the tree.

The raw p-value, also reported in each node, is 3.32E-10. Bonferroni adjustment multiplies this by the number of splits that are possible at each node. Note that the number of possible splits is less than 8189 because, among the 10000 selected compounds, some of the explanatory variables might either be constant or perfectly correlated. The adjustment used counts and removes perfectly correlated variables as well as variables that are constant zero or one.

The SCAM program allows to control the Minimum Split Size (MSS), which is defined as the minimum number of compounds required to be in each daughter node after the split. A low value of MSS can create splits that put too few compounds in one of the two daughter nodes and might thus focus on outliers rather than more general structural features. On the other hand, if the value is too high the search for a significant split might fail. We followed a general process of setting MSS relatively large at the beginning of the tree building process and progressively decreased MSS as the tree progressed. Splitting of a tree stops when MSS equal to two does not allow further splits of any of the current terminal nodes.

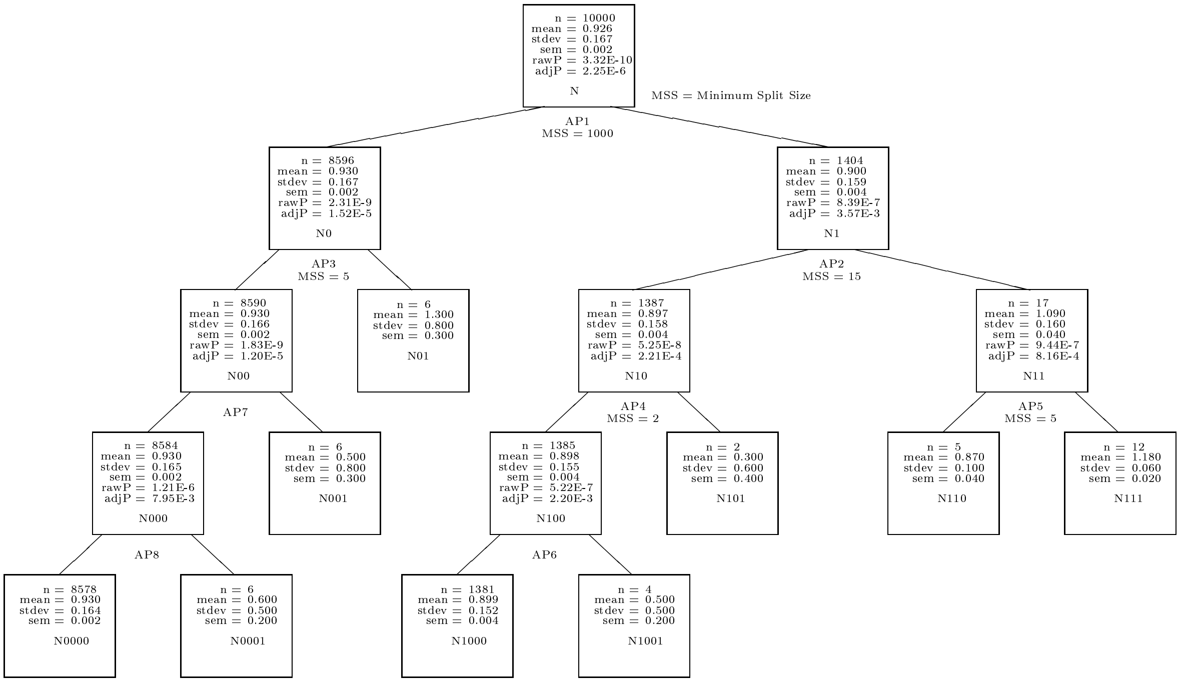

FIG. 3. *SCAM tree based on 10000 randomly selected compounds. Not shown are two further splits of node N0000 obtained by setting MSS equal to two. This leads to a final junk pool of size 8573.*

In the display of the tree, the number of compounds is reported for each node. Also given are the mean of the potencies in the node, the standard deviation, and the standard error of the mean. We are particularly interested in splits where the compounds in the node on the right (atom pair present) show a higher average than those in the left node (atom pair missing). These are called *positive rules*, because they identify atom pairs associated with (high) potency. As an example, the splits of nodes N0 and N11 in Figure 3 give positive rules. Similarly, the presence of an atom pair leading to a significantly lower average than its absence will be called a *negative rule*, see for example the splits of nodes N00 and N10 in Figure 3. The leftmost terminal node N0...0 is not defined by the presence of any atom pairs (no positive rules); we refer to it as the "junk" node.

Using the tree[2], any untested compound is predicted to belong to the terminal node determined by its atom pair description, and its potency is predicted to be the average potency in that terminal node. Presented with an untested compound, the tree will predict the potency. Additionally, the rules defining terminal nodes with high average potency suggest molecular features important for binding.

The t-test with Bonferroni adjustment replaces the cross-validation and pruning techniques used in CART and makes computation tractable. The often spectacularly small p-values encountered in this approach should not be taken too seriously. Our analysis is exploratory; we want to find good regions of the chemical space and do not want to be led astray too often. Exceedingly small p-values arise because of the large sample sizes and also arise when binding is governed by a relatively few sharp features.

The collection of 70223 compounds under consideration, like many such collections, does not cover a large part of "chemical space". Chemists often synthesize many compounds that are similar to a useful compound so there are likely to be substantial numbers of closely related compounds in the collection - a collection is more like a star cluster or galaxy than a uniform random set. Nonetheless, the methodology we use, in essence an exploratory device, has useful implications as we shall see below.

4. DESIGNING A SEQUENTIAL SCREENING SCHEME

The first stage of a sequential approach as depicted in Figure 1 requires specification of an initial sample size and a strategy to select the sample. Once done, and the potencies obtained, designing a second stage of sampling should exploit the information gathered in the first stage. This process can be continued over several cycles of selection.

---

[2] A tree as shown in Figure 3 can also be regarded as a (linear) regression tree by taking

$$\begin{aligned}\log(\text{potency}) \quad = \quad & \gamma_0 + \gamma_1(1-X_1)(1-X_3)(1-X_7)X_8 + \gamma_2(1-X_1)(1-X_3)X_7 + \gamma_3(1-X_1)X_3 \\ & + \gamma_4 X_1(1-X_2)(1-X_4)(1-X_6) + \gamma_5 X_1(1-X_2)(1-X_4)X_6 \\ & + \gamma_6 X_1(1-X_2)X_4 + \gamma_7 X_1 X_2(1-X_5) + \gamma_8 X_1 X_2 X_5 + \text{error}.\end{aligned}$$

Here X_k is the binary explanatory variable indicating whether the kth atom pair occurs in a compound or not, $k = 1, \ldots, 8$. The least squares estimates of $\gamma_0, \gamma_1, \ldots, \gamma_8$ are related to the average potencies in the terminal nodes from left to right.

A general and more encompassing sequential decision approach would specify appropriate loss functions, perhaps a prior distribution on the function describing the connection between structure and activity, and compute solutions. It is unclear to us whether a procedure is available that is computationally feasible for the problems of the scale presented here.

For the testbed problem described in Section 2 we will consider five factors that could play an essential role in defining a sequential strategy. To carry out an initial screening study to determine the most relevant of those factors, we will consider each of them at only two levels.

N1 Number of first stage samples

In determining an initial sample size on heuristic grounds we took into account two conditions. First, since the analysis based on the initial sample relies on constructing a tree, this tree should have several terminal nodes with positive rules in order to be useful. Second, practicing chemists believed that far more than 10000 (of the more than 70000) compounds would have to be tested. After careful consideration, we chose the two levels of N1 to be 5000 and 10000.

D1 Design of first stage

The design for the first stage sample can depend only upon the information about the chemical structures available. Given a distance on the space of compounds we could, in principle, select an optimum set following criteria and methods described in Johnson, Moore, and Ylvisaker (1990) or Haaland, McMillan, Nychka, and Welch (1994). Similarity indices such as the ones described in Finn (1996) could provide such measures of distance between compounds. But the computational effort required to obtain designs that optimize some criterion is beyond current capabilities for problems of the scale facing us. We therefore introduce two alternative strategies.

The first strategy rank orders the compounds by their Burden numbers. Then, starting with the compound with the largest Burden number, we successively choose every 7th compound until a sample of size 10000 is obtained. The design of size 5000 is obtained by selecting every other compound from the set of size 10000. We refer to this method as systematic sampling by Burden numbers, SSBN.

The second strategy uses clustering, ideally to form clusters of compounds similar within the cluster but dissimilar between clusters. Monothetic clustering as described in Kaufman and Rousseeuw (1990) seemed an appropiate tool for the binary atom pair descriptors at hand. However, different from their approach, we define the similarity between two compounds as the ratio of the number of atom pairs they have in common and the total number of atom pairs occurring in either of the two compounds. This index dates back to Jaccard (1908) but nowadays gains increasing popularity as the Tanimoto coefficient, see for example Van Drie and Lajiness (1998). It is used because of the asymmetry in the descriptors: For two molecules, having an atom pair in common is more informative than both of them having the same atom pair missing. We refer to this clustering on atom pairs by CLAP.

Approximately the same number of compounds was selected from each cluster to obtain a starting set of 10000. The selection within each cluster was made based on the rank ordering by Burden numbers as described above. A first stage design of size 5000 was obtained by choosing every other compound within each cluster.

ST Number of stages

In the normal course of business it is impractical to employ a fully sequential procedure or even one that requires more than two or three stages (counting the initial stage) of selection and assay. Therefore, the two levels of ST are 2 and 3.

N2 Sample size at additional stages

An arbitrary decision was made to restrict attention to procedures that either take a total of 2500 or 5000 new samples after the first stage with equal numbers at both stages 2 and 3 if ST=3. Thus, a three stage procedure with 2500 new samples means that 1250 samples are taken at each of the two additional stages.

D2 Design of additional stages

Which compounds to select at the subsequent stages needs attention. We start by defining a good node as one where the average potency (the observed average from the tested compounds of earlier stages ending up in this node) is greater than 1.05. This value is chosen rather arbitrarily and approximately corresponds to the upper empirical 15% point of the data. There are two types of nodes remaining: the poor nodes and the junk node. All untested compounds are classified (predicted) to lie in one of the terminal nodes.

It appears reasonable to select the second stage sample from those compounds predicted to lie in the good nodes. However, there may be insufficient numbers of such compounds. Moreover, there may be many good compounds in the other nodes, particularly in the junk pool. Accordingly, we decided to compare two different strategies. In one strategy we select (if possible) 90 percent of the additional compounds from good nodes of the previously constructed tree and the other 10 percent from the remaining nodes (90/10). The second strategy aims to select equal numbers from the

Table 2

Five factors characterizing a sequential screening scheme. Each factor is studied at two levels.

Factor	Symbol	Levels
Number of first stage samples	N1	5000, 10000
Design of first stage	D1	SSBN, CLAP
Number of stages	ST	2, 3
Sample size at additional stages	N2	2500, 5000
Design of additional stages	D2	90/10, 50/50

good nodes as from the remaining nodes (50/50). More explicitly:

(I) *Good node selection*: Start with the good node with highest average potency. Compounds predicted to be in this node are chosen until 90 (or 50) percent of the $N2/(ST - 1)$ additional samples are found. If there are too few compounds predicted to be in this node, proceed to the good node with the next largest average. If there are too few compounds found in the good nodes to date, continue sampling from the terminal node (but not the junk node) with the next largest average potency.

(II) *Similarity selection*: The remaining 10 (or 50) percent of the $N2/(ST - 1)$ additional samples are selected from terminal nodes not sampled in (I). To do so, all (tested) compounds from previous stages that fall in these nodes are rank ordered by potency. Then, starting with the most potent of these compounds, we take the five nearest neighbors based on Burden numbers and continue until the desired number of $N2/(ST - 1)$ additional samples is reached. The anticipation is that chemical structures are added that have atom pair features potentially leading to positive rules in the next round of recursive partitioning; it is similar to the practice of testing new compounds with substructures that are similar to active compounds.

A summary of all five factors is provided in Table 2.

Variations on the rules described in (I) and (II) can be explored. For example, instead of solely using the average potency in a node, the variability might be taken into account as well. We will not pursue this point.

5. EXPERIMENTAL LAYOUT AND EVALUATION CRITERIA

In order to investigate the effect of the five factors in Table 2, we could consider all $2^5 = 32$ possible combinations and evaluate the performance of each of the resulting screening strategies. This would allow the estimation of the main effect of each factor as well as all higher order interactions. To reduce the computational effort and, on the premise that interaction effects of order higher than two may be negligible, we use a half fraction of the complete 2^5 design (Box and Draper, 1987, page 148), leading to 16 different screening strategies. These are shown in the left part of Table 3. This design allows the identification of all main effects and all two factor interactions.

Each of the 16 rows in Table 3 fully describes a screening strategy. Recursive partitioning is applied to the total of N compounds assayed. The resulting tree is then used to predict the activity of the remaining $70223 - N$ compounds. Of those, the molecules predicted to be in good nodes will be screened.

As mentioned earlier, lead compounds identified in screening campaigns generally need further structural modifications to improve their biological and chemical properties. To do so, a medicinal chemist typically starts modifying a compound by exchanging different functional groups of the molecule. This approach is quite time consuming and thus only a few different and the most promising leads resulting from a screening campaign can be considered. Due to these considerations, our strategy focused on identifying the best 100 compounds in a given collection. We refer to these as the top100 compounds and the goal is to identify as many of these as possible. The potencies of the top100 compounds for the present data set range from 1.682 to 3.102 on the logged scale.

From Table 3, note that the total number N_T of compounds tested for each of the 16 runs varies, as it is given by the sum of N and the additional compounds among the $70223 - N$ that are predicted to be in good nodes of the final tree. We therefore compare the actual number of top100 compounds found by each of the 16 strategies to the expected number of top100 compounds we would find by randomly

TABLE 3

A 2_5^{5-1} design for five factors leading to 16 different screening strategies and the total number N of compounds that need testing under each strategy. The last column gives the results for the evaluation criterion.

Strategy	N1	D1	ST	N2	D2	N	I_{100}
1	10000	CLAP	3	2500	90/10	12500	1.11
2	5000	CLAP	3	2500	50/50	7500	2.06
3	5000	CLAP	3	5000	90/10	10000	1.81
4	10000	CLAP	3	5000	50/50	15000	1.69
5	5000	SSBN	3	2500	90/10	7500	0.66
6	10000	SSBN	3	2500	50/50	12500	1.07
7	10000	SSBN	3	5000	90/10	15000	1.02
8	5000	SSBN	3	5000	50/50	10000	1.31
9	5000	CLAP	2	2500	90/10	7500	1.26
10	10000	CLAP	2	2500	50/50	12500	1.85
11	10000	CLAP	2	5000	90/10	15000	1.31
12	5000	CLAP	2	5000	50/50	10000	1.71
13	10000	SSBN	2	2500	90/10	12500	0.83
14	5000	SSBN	2	2500	50/50	7500	1.29
15	5000	SSBN	2	5000	90/10	10000	1.33
16	10000	SSBN	2	5000	50/50	15000	1.71

selecting N_T molecules among the 70223. More formally, we define

$$I_{100} = \frac{\text{number of top100 compounds found by systematic screening}}{\text{expected number of top100 compounds found by random screening}}$$

as being the improvement of a systematic screening strategy over random sampling. This is the quantity reported in the last column of Table 3. Section 6 presents the analysis of these results.

6. EXPERIMENTAL ANALYSIS

To gain an initial impression of the most important effects, we did an analysis of variance on I_{100}. All main effects and all two factor interactions were included. Figure 4 shows a half normal probability plot of the resulting effects. Two factors, the design of the first stage (D1) and the design of the additional stages (D2) appear to be most important.

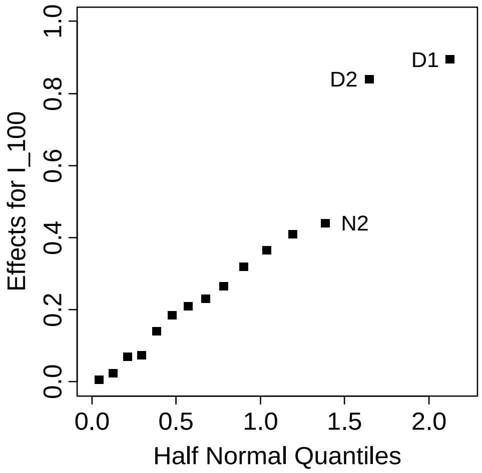

Fig. 4. *Half normal probability plot of the main effects and two-factor interaction effects for* I_{100}.

In Figure 5, boxplots for the two levels of each factor display the main effects. Each boxplot is based on eight values. Figure 5 supports the effects found for D1 and D2. Clustering on atom pairs (CLAP) appears superior to systematic sampling by Burden numbers (SSBN). This seems plausible, as atom pairs provide more detailed information on the chemcial structure of a compound than the univariate Burden number. The result could be compound selections that are more representative of the entire collection, which in turn leads to a better SCAM model.

The 90/10 split at the second stage design seems less effective than the 50/50 split. A three stage procedure does not appear more effective than a two stage procedure. For later studies we therefore elect to not go beyond two stages and we also adopt the 50/50 split for D2.

Conclusions about N1 and N2 are unclear. Whether N1 is 5000 or 10000 does not seem to matter much. The effect of N2 is ambiguous but leans to the choice of N2=5000. As expected from Figure 4, interaction plots (not shown here) did not show evidence of strong interactions among the five factors.

Some shortcomings of our analysis have led to additional experiments aiming at confirming the above findings. These are the subject of Section 7 below.

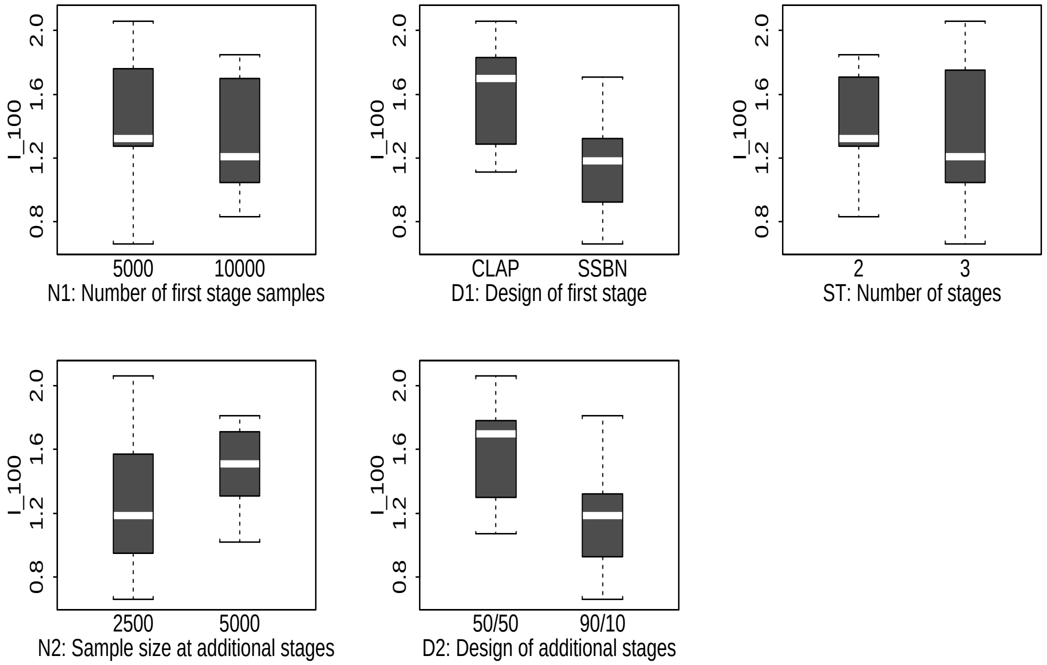

FIG. 5. *Main effects of each factor for* I_{100}.

7. ADDITIONAL EXPERIMENTATION

Our feeling for the need of additional experimentation primarily arose from the fact that all results reported in Table 3 are essentially based on two initial compound selections of size 10000, chosen according to SSBN and CLAP. This leads to dependence among the reported values for I_{100} and does also not reflect the variability that might result from the choice of different initial designs. The results based on analysis of variance might thus be questionable. Additional exploratory analyses were therefore carried out and are the subject of this section. The main findings can be summarized as follows.

1. The method chosen for the selection of the initial design appears unimportant (see Subsection 7.1). Systematic approaches show no benefits over a random selection. The effect of D1 shown in Figures 3 and 4 is thus spurious.

2. No general recommendation is possible on the size of the initial design (see Subsection 7.2). Very small samples can suffice at the first stage if they contain structure-activity information.
3. Sequential sampling is beneficial compared to applying recursive partitioning to a one-step selection (see Subsection 7.3).

7.1 Initial sample design

Figure 5 above suggests evidence for a strong effect due to the choice of the initial design scheme. This effect might be spurious, as the results in Table 3 are based on fixed initial designs of sizes 5000 and 10000 for each of CLAP and SSBN. Possible variability in the response could thus arise if the starting points used for selecting those designs are varied. In particular, for D1=SSBN, we could pick every 7th compound beginning with the molecule having the 2nd largest rather than the largest Burden number.

To fully replicate the 16 runs several times with varied SSBNs and CLAPs was computationally prohibitive because of the logistical complexity of the experiment. Instead we experimented by fixing N1=10000, ST=2, N2=5000, and D2=50/50. We then generated four different SSBN designs, four different CLAP designs and four independent random designs (RAND). The four SSBN designs (see Section 4) were generated by picking every 7th compound beginning with the molecule having the kth largest Burden number, $k = 1, 2, 3, 4$. Similarly, to select the four CLAP designs, we simply changed the starting compound for the systematic Burden number sampling within each cluster without changing the underlying clustering of the 70223 compounds. Each of the twelve samples is used as a starting design (D1) for our sequential screening method; the results obtained for I_{100} are summarized in Table 4.

All three methods produce similar means and substantial variability. The smaller variability associated with CLAP does not overcome its computational disadvantages. The most surprising result is that random selection does about as well as the procedures using the chemical structures of the compounds. One reason may be that the designs all try to cover a very high (over 8000) dimensional space and none can do so very effectively, see Young, Farmen, and Rusinko (1996). The near equivalence of

TABLE 4

Response I_{100} *for three different methods of selecting the initial sample, each replicated four times.*

Sample	SSBN	CLAP	RAND
1	1.17	1.17	0.97
2	1.51	1.33	1.34
3	1.71	1.48	1.79
4	1.87	1.69	2.00
mean	1.57	1.42	1.52
stdv	0.30	0.22	0.46

these three first-stage design schemes has been borne out when using different assays and other sets of compounds as well, see Section 8.

7.2 Initial sample size

Figure 5 indicates little effect on I_{100} from changing the size of the initial sample from 5000 to 10000. To explore this further we fix the total sample size N at 15000, and let the initial sample size vary (N1=5000, 7500, 10000). The levels for the other factors were set at D1=SSBN, ST=2, and D2=50/50. Again, four repeated samples were taken as in Table 4; the results are summarized in Table 5.

As the results in Table 5 show, increasing the number of initial compounds selected does not necessarily improve the overall hit ratios. The reason is that even a large sample might provide little information about the relationship between chemical structure and activity and thus lead to a poor selection of compounds for the later stages. For N1=10000, closer inspection of the trees showed that among the four initial SCAM trees three were “good” in the sense of having at least three good terminal nodes (see the discussion of D2 for the definition of a “good” terminal node). Only two good initial SCAM trees turned up when N1=5000 and only one when N1=7500. This is what is reflected in the results of Table 5.

A conclusion that can be drawn from this experiment is that the initial sample size should be large enough to produce a tree with an adequate number, three or more apparently, of good terminal nodes. A possible approach would be to select an

TABLE 5

Response I_{100} *for three different initial sample sizes while keeping the total sample size fixed.*

Sample	N1=10000	N1=7500	N1=5000
1	1.17	0.97	1.18
2	1.51	1.06	1.23
3	1.71	1.08	1.42
4	1.87	1.32	1.46
mean	1.57	1.11	1.32
stdv	0.30	0.15	0.14

initial sample of size 2500 say, build a tree and examine its adequacy. Take another sample of 2500 if the tree is inadequate. This runs against the obstacle of setup costs for each stage but is essential because going ahead with an inadequate tree will be of little utility.

Table 5 exhibits a decrease in variability as the initial sample size decreases. Since the total sample size N is fixed, the second stage sample size increases as the initial sample size decreases. Because the second stage sample is expected to be more homogeneous than the initial sample, a decrease in variability should be expected.

7.3 Benefits from sequential sampling

Is there a benefit from the sequential strategy? Starting with the compounds having the 1st, 2nd, and 3rd largest Burden number, we systematically sampled 15000 compounds. The average values and standard deviations obtained for I_{100} (1.05 ± 0.19) and $I_{350}(1.00 \pm 0.14)$ are significantly worse than the average of the two corresponding columns headed SSBN in Table 4. There appear to be real benefits from using a sequential scheme

8. CONFIRMATION EXPERIMENTS

Two experiments are used to validate the findings of the proposed precedures. In the first experiment, using the same set of 70223 compounds and the methods developed and analyzed above, a second assay was explored to confirm the effectiveness of the approach as well as the earlier conclusion that neither the initial sample design nor the initial sample size play an important role. The factor ST was fixed at two and N2 was set at 5000. For D2 we used (I) and (II) (see Section 4) together with a 50/50 split. In the second experiment, a different set of 52883 compounds, each of which was tested in two different assays, are studied.

In the first confirmation experiment, the combinations of the levels of N1 and D1 produce four runs. We also included a run with the initial design being a random sample of size 5000. Each of the five strategies was then replicated three times. The results are shown in Table 6. For the first four runs, the replicates were produced similar to those in Table 4. Note that strategies 2 and 3 in Table 6 correspond to strategies 12 and 16 in Table 3.

Table 6

A confirmation experiment exploring N1 and D1 based on the same 70223 compounds tested in a different assay.

Strategy	N1	D1	I_{100}			mean	stdev
1	10000	CLAP	1.85	1.95	2.27	2.02	0.22
2	5000	CLAP	1.90	2.16	2.18	2.08	0.16
3	10000	SSBN	1.93	1.94	2.17	2.01	0.14
4	5000	SSBN	1.63	2.08	2.28	2.00	0.33
5	5000	RAND	1.66	1.71	2.57	1.98	0.51

Considering the three replicates for each run in Table 6 as independent, an analysis of variance of the first four rows of the data reveals no significant effects. Again, the surprising fact that RAND, while more variable, appears to be as good as CLAP or SSBN for D1. For this assay, the choice of N1=5000 is as productive as choosing N1=10000, reflecting the fact that N1=5000 already produced an adequate tree with three good terminal nodes.

For further verification, we studied a set of 52883 compounds with two different assays. We chose ST=2. Implementing the approach suggested in Section 7.2, we used N1=2500 and found that no further augmentation was necessary. We chose N2=N1, dropped CLAP and compared SSBN and RAND. For D2 we used (Ia) and (II) together with a 50/50 split, where (Ia) is a modification of (I) [3].

(Ia) Select compounds among those that are predicted to be in the node with the highest average potency until 50 (or 90) percent of the N2/(ST − 1) additional samples are found. If the number of compounds in this node is not sufficient, go to the node with the next highest potency provided its average potency is at least as large as the average potency of the sample used to construct the tree. Continue as long as possible; otherwise go to step (II).

Each of the two strategies was repeated four times. The resulting ratios I_{100} are shown in Table 7.

TABLE 7

Response I_{100} *for a second set of 52883 compounds screened in two different assays.*

Sample	Assay 1		Assay 2	
	SSBN	RAND	SSBN	RAND
1	4.16	4.00	2.07	1.91
2	4.92	5.21	3.43	1.83
3	4.42	2.77	2.82	2.33
4	4.10	3.80	1.63	2.59
mean	4.40	3.95	2.49	2.17
stdev	0.37	1.00	0.80	0.36

Although different for the two assays, the results again demonstrate the benefits from the sequential sampling scheme. Compounds are often screened in multiple assays to explore different biological properties. A given set of descriptors might not be equally effective in capturing the relevant chemical structures leading to favorable responses in all assays. More importantly, the relative assay variability can vary among assays affecting the SAR and the rate at which desirable compounds will be detected. As Table 7 again shows, a systematically selected initial sample does not lead to substantive improvements over a random selection.

The hit ratios considered so far evaluate the performance of the entire screening

---

[3] The modification is to accommodate the case when there may be very few compounds predicted to be in good terminal nodes of the first stage tree.

strategy. For the 52883 compounds data sets, we also considered a different criterion for comparing performance. The goal is to evaluate the gain achieved *after* the initial sample, thus focusing on the ability of SCAM to direct the search towards potent areas of the chemical space. In the first stage sample of size 2500 selected by SSBN we found 5 of the top100 compounds. With N2=2500 a final tree is built and of the remaining 47883 molecules 728 are predicted to be in good nodes. These are subjected to screening and among the total of 3228 compounds assayed after the first stage, 40 were found to be among the top100. Therefore, using SCAM to preselect compounds for assaying, top100 compounds turn up at an average rate of 40/3228=0.0124 or, in other words, one out of about 80 compounds assayed is among the top100. Without the initial SCAM tree to virtually prescreen the 52883-2500=50383 compounds, we might assay all of them to find all the remaining 100-5=95 top100 compounds. On average, top100 compounds would be discovered at a rate of 95/50383=0.0019, which corresponds to one molecule out of about 526. The SCAM gain rate relative to random sampling is now defined as (40/3228)/(95/50383)=6.57. Table 8 summarizes the SCAM gain rates G_{100} for both assays. This indicates that SCAM can rapidly and efficiently guide the process of compound selection to active regions of the chemical space. This will be especially useful, we believe, in dealing with virtual libraries.

TABLE 8

SCAM gain rate G_{100}. *The values are arranged in correspondence to those in Table 7.*

Sample	Assay 1		Assay 2	
	SSBN	RAND	SSBN	RAND
1	6.57	5.73	3.23	2.51
2	7.61	8.16	5.51	2.18
3	6.94	4.21	4.04	3.35
4	6.27	5.96	2.42	4.27
mean	6.85	6.01	3.80	3.08
stdev	0.58	1.63	1.32	0.94

9. OTHER DIRECTIONS

Several issues need fuller exploration. One is the use of multiple trees in place of the "greedy" single tree used above. Recently Tatsuoka, Gu, Sacks, and Young (2000) introduced predictors based on multiple trees, tailored for accurate prediction of extreme values. Variabilites of hit ratios from different initial samples might come from the lack of stability of a SCAM tree and the difference in the number of additional compounds predicted to be in good nodes. Tatsuoka et al. (2000) note that predictors based on multiple trees are more accurate and less variable than the single SCAM tree. To reduce the variabilities, a sequential strategy could be used in conjunction with multiple tree predictors; such a study is now underway.

A second major question is connected with the implications of measurement errors in the assay. These errors ought to be incorporated into the formulation of objective functions for comparing strategies. Practice thus far indicates that, even without taking the errors into account, the sequential strategies are effective and are currently in use at GlaxoWellcome (Jones-Hertzog et al., 2000).

A third issue to be addressed is the effect of scaling up: treating hundreds of thousands of compounds, not "merely" 70000. Combinatorial chemistry (Service, 1996) is one arena where such scales (and greater ones) will be present. Combinatorial schemes allow the electronic generation of databases of compounds by considering all combinations of a given group of molecular building blocks. Because synthesis of molecules is not cheap (it is even more expensive than typical assays), new questions will arise here if we take the cost of synthesis into account.

In practice, biological activity is not the only quantity of interest. The same compounds are commonly tested in several assays to determine other biological properties such as toxicity. Sequential screening schemes that allow the handling of multivariate measures are currently under investigation.

A modified version of recursive partitioning allowing for the extraction of multiple chemical features at each node has recently been published by Cho, Shen, and Hermsmeier (2000). Gobbi, Poppinger, and Rohde (1997) used a genetic algorithm to identify lead compounds. Although their approach certainly also has the ability to identify good starting points for future optimization by medicinal chemists, a disadvantage is that it does not clearly pinpoint the relevant structural features. Friedman and Fisher (1999) discussed a new algorithm for identifying regions where some response is maximized over a high dimensional space. Their approach can be seen as a generalization of recursive partitioning, as it divides the search space in more general types of "boxes". The method appears effective but has, to the best of our knowledge, not yet been applied to problems in the area of drug discovery.

Different statistical techniques for modeling structure-activity relationships are used at the later "compound optimization" stage of the drug development process, where the medicinal chemists systematically modify hits resulting from initial screening campaigns in order to improve their biological properties. The number of compounds to be dealt with might be in the hundreds only and the molecules are generally also more homogeneous in terms of their chemical structure. The most frequently used statistical modeling tools at this point are Regression Analysis (Patankar and Jurs, 2000), Partial Least Squares (PLS; Helland, 1990), Neural Networks (Kauffman and Jurs, 2000), or combinations thereof (Viswanadhan, Mueller, Basak, and Weinstein, 1996). Many variants of these have been developed and tuned to the needs of the chemists. The recent conference proceedings of the 12-th European Symposium on Quantitative Structure-Activity Relationships (Gundertofte and Jørgensen, 2000) cover applications of all of these.

10. SUMMARY AND CONCLUDING REMARKS

Exhaustive screening of libraries and other large sets of chemical compounds is not uncommon for finding good lead compounds in a drug discovery process. Despite the automation of the processes of synthesizing and assaying compounds, inefficiencies and costs can become prohibitive. Sequential screening strategies are potentially valuable for finding potent lead compounds while controlling costs. A class of procedures studied here combines simple chemical descriptors of molecules, recursive partitioning and careful computational algorithms to produce ad hoc sequential designs that are effective. The potential merits of such tactics are now receiving some attention (Walters, Stahl, and Murko, 1998). What we have presented here are studies of how such methods can be implemented and questions that should be addressed. This is an arena where statistical insight can be influential and one that generates a variety of unexplored, interesting problems.

Due to proprietary rights, the data sets used in this work can unfortunately not be made available for public use. However, a data set containing activity data and structural information of over 30000 compounds is available from the homepage of the National Cancer Institute at http://dtp.nci.nih.gov.

ACKNOWLEDGMENTS

We are grateful to Andy Rusinko for computing the atom pair descriptors and the Burden numbers for the two sets of compounds used in this work. He also provided very efficient software to rapidly predict untested compounds via a SCAM tree. Our thanks are also to David Cummins and Scott Langfeldt for their contribution in numerous discussions and for preparing text file versions of the atom pair information which we used at various stages of this work. We also gratefully acknowledge valuable input of Alan Menius and William Welch that helped to shape the current form of the sequential search scheme. The editor as well as an anonymous referee provided very helpful comments that greatly improved the presentation of the paper.

REFERENCES

Box, G. E. P. and Draper, N. R. (1987). *Empirical model-building and response surfaces.* Wiley Interscience, New York.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and regression trees.* Wadsworth, Belmont.

Burbaum, J. J. (1998). Miniaturization technologies in HTS: How fast, how small, how soon? *Drug Discov. Today* **3** 313-322.

Burden, F. R. (1989). Molecular identification number for substructure searches. *J. Chem. Inf. Comput. Sci.* **29** 225-227.

Carhart, R. E., Smith, D. H., and Venkataraghavan, R. (1985). Atom pairs as molecular features in structure-activity studies: Definition and applications. *J. Chem. Inf. Comput. Sci.* **25** 64-73.

Cho, S. J., Shen, C. F., and Hermsmeier, M. A. (2000). Binary formal inference-based recursive modeling using multiple atom and physiochemical property class pair and torsion descriptors as decision criteria. *J. Chem. Inf. Comput. Sci.* **40** 668-680.

Cortese, R. (Ed.) (1996). *Combinatorial libraries. Synthesis, screening and application potential.* De Gruyter, Berlin.

Finn, P. W. (1996). Computer-based screening of compound databases for the identification of novel leads. *Drug Discov. Today* **1** 363-370.

Friedman, J. H. and Fisher, N. I. (1999). Bump hunting in high-dimensional data. *Statistics and Computing* **9** 123-143 (with discussion).

Gobbi, A., Poppinger, D., and Rohde, B. (1997). Finding biological active compounds in large databases. Internet publication,
available at http://www.unibasel.ch/mdpi/ecsoc/f0008/f0008.htm.

Gundertofte, K. and Jørgensen, F. S. (Eds.) (2000). Molecular modeling and prediction of bioactivity. Procedings of the 12-th European Symposium on Quantitative Structure Activity Relationships. Kluwer Academic/Plenum Publishers, New York.

Haaland, P. D., McMillan, N. J., Nychka, D. W., and Welch, W. J. (1994). Analysis of space filling designs. *Computing Science and Statistics* **26** 111-120.

Hawkins, D. M. (1994). *FIRM Formal Inference-based Recursive Modeling.* Release 2, University of Minnesota, St. Paul, MN.

Hawkins, D. M. and Kass, G. V. (1982). Automatic interaction detection. In *Topics in Applied Multivariate Analysis*, Hawkins, D.M., Ed. 269-302. Cambridge University Press, Cambridge, UK.

Helland, I. S. (1990). Partial least squares regression and statistical models. *Scand. J. Statist.* **17** 97-114.

Jaccard, P. (1908). Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaud. Sci. Nat.* **44** 223-270.

JOHNSON, M. E., MOORE, L. M., and YLVISAKER, D. (1990). Minimax and maximin distance designs. *J. Statist. Plann. Inference* **26** 131-148.

JONES-HERTZOG, D. K., MUKHOPADHYAY, P., KEEFER, C. E., and YOUNG, S. S. (2000). Use of recursive partitioning in the sequential screening of G-protein-coupled receptors. *J. Pharmacol. Toxicol.* **10** 207-215.

KAUFFMAN, G.W. and JURS, P.C. (2000). Prediction of inhibition of the sodium ion-proton antiporter by benzoylguanidine derivatives from molecular structure. *J. Chem. Inf. Comput. Sci.* **40** 753-761.

KAUFMAN, L. and ROUSSEEUW, P. J. (1990). *Finding groups in data.* Wiley Interscience, New York.

PATANKAR, S.J. and JURS, P.C. (2000). Prediction of IC50 values for ACAT inhibitors from molecular structure. *J. Chem. Inf. Comput. Sci.* **40** 706-723.

RUSINKO, A. III, FARMEN, M. W., LAMBERT, C. G., BROWN, P. L., and YOUNG, S. S. (1999). Analysis of a large structure/biological activity data set using recursive partitioning. *J. Chem. Inf. Comput. Sci.* **39** 1017-1026.

SERVICE, R. F. (1996). Combinatorial chemistry hits the drug market. *Science* **272** 1266-1268.

TATSUOKA, K., GU, C., SACKS, J., and YOUNG, S. S. (2000). *Predicting extreme values in large datasets. J. Comput. Graph. Statist.* To appear.

VAN DRIE, J. H. and LAJINESS, M. S. (1998). Approaches to virtual library design. *Drug Discov. Today* **3**, 274-283.

VISWANADHAN, V. N., MUELLER, G. A., BASAK, S. C., and WEINSTEIN, J. N. (1996). A new QSAR algorithm combining principal component analysis with a neural network: Application to calcium channel antagonists. Internet publication, available at http://org.chem.msu.su/people/baskin/neurchem.html.

WALTERS, W. P., STAHL, M. T., and MURKO, M. A. (1998). Virtual screening - an overview. *Drug Discov. Today* **3** 160-178.

WESTFALL, P. H. and YOUNG, S. S. (1993). Resampling-based multiple testing: Examples and methods for *p*-value adjustment. Wiley Interscience, New York.

Young, S. S., Farmen, M. W., and Rusinko, A. III (1996). Random versus rational: Which is better for general compound screening? *Network Science*, Internet publication,
available at http://www.netsci.org/Science/Screening/feature09.html.