Exploration of Cluster Structure−Activity Relationship Analysis in Efficient High-Throughput Screening

X. Sunny Wang,*,[†] G. A. Salloum,[‡] H. A. Chipman,[§] William J. Welch,[||] and S. Stanley Young[⊥]

Department of Statistics and Actuarial Science, University of Waterloo, Ontario, Canada, Department of Mathematics, Camosun College, British Columbia, Canada, Department of Mathematics and Statistics, Acadia University, Nova Scotia, Canada, Department of Statistics, University of British Columbia, British Columbia, Canada, and National Institute of Statistical Sciences, 19 T.W. Alexander Drive, Research Triangle Park, North Carolina 27709-4006

Received October 26, 2006

Sequential screening has become increasingly popular in drug discovery. It iteratively builds quantitative structure−activity relationship (QSAR) models from successive high-throughput screens, making screening more effective and efficient. We compare cluster structure−activity relationship analysis (CSARA) as a QSAR method with recursive partitioning (RP), by designing three strategies for sequential collection and analysis of screening data. Various descriptor sets are used in the QSAR models to characterize chemical structure, including high-dimensional sets and some that by design have many variables not related to activity. The results show that CSARA outperforms RP. We also extend the CSARA method to deal with a continuous assay measurement.

1. INTRODUCTION

In drug discovery, pharmaceutical companies have widely adopted high-throughput screening (HTS) of large compound libraries as a method to identify drug candidates.[1] However, HTS alone cannot fulfill the job of screening all compounds, as the estimated number of possible drug molecules is roughly 10^{40}.[2] Hence sequential screening[3,4] has been developed to help reduce costs and make HTS more efficient.

Sequential screening combines HTS and virtual screening (a screening model) in one integrated screening process. Instead of testing an entire chemical library against a biological target, only a fraction of the library, known as the initial sample set, is assayed. The purpose of the initial sample set is not to find as many actives as possible but simply to collect data on a diverse set of compounds in the chemical space, so that a computational analysis of these data can identify trends and help to select a further set of compounds to be screened.

The sequential screening process is illustrated in Figure 1. The work flow is as follows: (1) The experiment starts with the initial sample of compounds, which is run through the HTS process. (2) A quantitative structure−activity relationship (QSAR) analysis of the data is performed to identify structural, physiochemical, or other descriptors relevant to the biological activity. In this step, a model capable of predicting activity using chemical descriptor values is fit using the data from the initial set screened. (3) On the basis of the predictions from the first QSAR, the whole data inventory is virtually screened. The fitted model is used to predict activity for the untested compounds, in order to get a more focused set of compounds, i.e., the compounds predicted to be active. The chosen molecules are assayed in a second round of HTS, and the process iterates.

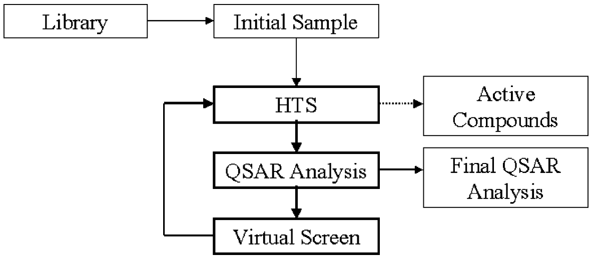

Figure 1. The sequential screening paradigm.

In this paper, a sequential screening method, cluster structure−activity relationship analysis (CSARA), is explored and compared to another popular QSAR method, recursive partitioning (RP). CSARA and RP are outlined in section 2. Empirical comparisons between CSARA and RP on low-dimensional descriptor sets are described in sections 3 and 4, while comparisons on high-dimensional sets are made in section 5. CSARA is extended to a continuous assay measurement variable in section 6 and briefly evaluated. Conclusions are drawn in section 7.

2. QSAR APPROACHES: CSARA AND RP

In theory, chemical compounds with similar structures will react with a biological target in a similar way.[5] Further, if the compounds with similar values of critical chemical descriptors can be grouped into one cluster, information about the activity of all compounds in the cluster may be obtained by simply assaying one or a few compounds randomly picked from the cluster. Engels and Venkatarangan[4] suggested such a cluster-based approach for sequential screening experi-

* Corresponding author e-mail: x29wang@math.uwaterloo.ca.
[†] University of Waterloo.
[‡] Camosun College.
[§] Acadia University.
[||] University of British Columbia.
[⊥] National Institute of Statistical Sciences.

10.1021/ci600458n CCC: $37.00 © 2007 American Chemical Society
Published on Web 05/05/2007

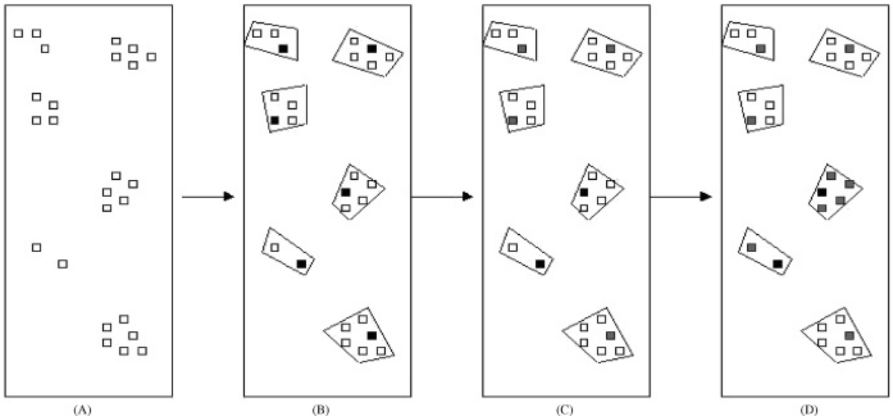

Figure 2. The CSARA process. Adapted from ref 4.

ments, namely cluster structure−activity relationship analysis (CSARA). Even though Engels and Venkatarangan[4] used several examples to show how useful the CSARA method is in the process of active compound selection compared to random selection, they did not compare CSARA with other QSAR methods, like recursive partioning (RP). As RP is a very popular QSAR approach in drug discovery, it provides a good benchmark for comparison. In this paper, we follow the same basic CSARA algorithm, but our focus is to deepen the understanding of CSARA and explore its efficiency for screening drug data relative to RP.

2.1. CSARA. The CSARA procedure is illustrated in Figure 2 and described as an algorithm below.

CSARA Algorithm

A. The entire compound library is available for sequential screening.

B. Compounds are clustered and one compound randomly selected from each cluster (shown in black in Figure 2 (B)) is assayed.

C. The randomly selected compounds, one from each cluster, are tested for biological activity. (In Figure 2 (C), the compounds shown in black are active and those in gray are inactive.)

D. All the compounds in the clusters with an active compound from (C) are assayed to get more accurate estimation of their biological activities.

The CSARA algorithm is based on the belief that compounds with similar chemical structures react with targets in a similar way. In Figure 2, CSARA partitions the entire compound library into six clusters, and one representative of each cluster is randomly selected and tested. If the selected compound is active, we place all other compounds belonging to that cluster in the focused library for the second round of HTS. However, if the selected compound is inactive, this would suggest that the remaining compounds in the cluster are also inactive, and they are not included in the second HTS. Although the illustration in Figure 2 has only six clusters, CSARA would typically use 100's or 1000's of clusters.

In fact, CSARA is a two-stage sequential screening process. Steps B, C, and D in the CSARA algorithm correspond to the key boxes in Figure 1. CSARA's step B is selecting an initial sample of compounds or training data. Step C is HTS to measure the activities for the initial sample. Step D combines QSAR analysis and virtual screening in order to get the focused library.

A critical part of CSARA is the cluster analysis. Partitioning the entire compound library into clusters can be approached with a wide variety of clustering algorithms.[6] Available methods of seeking clusters can be categorized broadly as hierarchical methods and partitioning methods. In this paper, we use one popular partitioning method, K-means.[7]

For the K-means algorithm, the user specified parameter K is the number of clusters. The K-means algorithm starts by randomly choosing K unique compounds as centers and then allocates all compounds to the cluster with the closest center. Each of the K cluster centers is then updated using the observations assigned to it. This algorithm iteratively alternates between allocating and recomputing until the cluster centers do not change with an update. Compared to other clustering algorithms, K-means is a fast algorithm, hence it is an appealing choice when dealing with large data sets, especially those arising from HTS. In the HTS context, K will determine the number of compounds assayed in the initial sample, as one compound is selected from each cluster.

The problem of partitioning data into K clusters is complex, and algorithms such as K-means can get stuck in local optima. A more effective implementation of K-means[8] with multiple restarts is used here to find good local optima.

2.2. Recursive Partitioning. Recursive Partitioning (RP) uses a tree-structured set of questions about the descriptor variables to recursively divide the data into groups in which the response variable is as homogeneous as possible. To build an RP model, the input space is recursively subdivided into nodes of a tree. To identify the best split for a specific node, the algorithm considers all possible binary splits for each descriptor variable and chooses the optimal one by some criterion.[9] This splitting is carried out recursively until some stopping condition is reached. Stopping criteria can be employed directly to choose tree size,[10] or a large tree can be grown and then pruned.[9]

Our research initially considered a pruning approach. When pruning a tree, a variety of criteria, including misclassification rate, the Gini index, and entropy, can be employed. For unbalanced classification problems such as drug discovery, where one class (i.e., active compounds) is rare, misclassification rate is an inappropriate pruning criterion. This is because we desire a tree that can rank compounds by the probability that they will be active, rather than just classify them as active/inactive. In our experiments, we used the Gini index for tree pruning.

In experiments with the AIDS antiviral data with 64 BCUT descriptors (described later in section 3.2), we made two

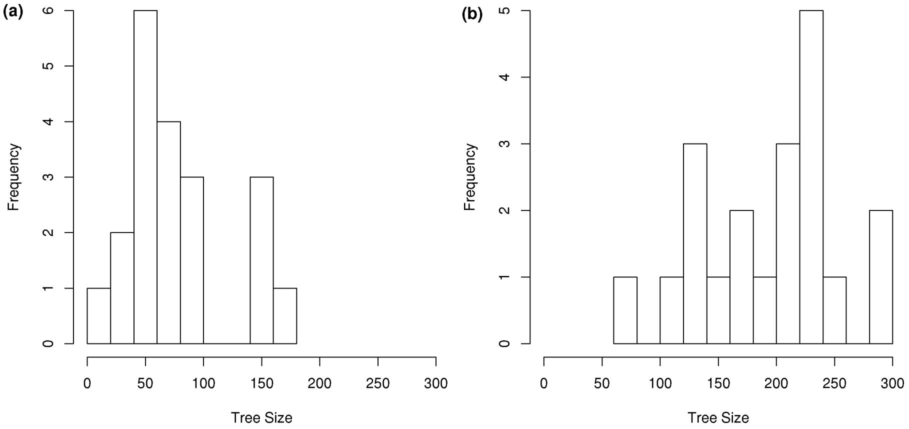

Figure 3. Histograms of the tree sizes chosen by (a) cross-validation and (b) performance on an independent test set.

unexpected discoveries with respect to pruning and tree size: 1. Cross-validation, the most common method of selecting tree size, appears to select a tree with too few nodes in unbalanced-class problems. 2. The largest trees yield the best (or near-optimal) out-of-sample predictive accuracy. In the remainder of this section, we outline the ideas behind these findings. Additional details are presented in the Appendix. A consequence is that for the remainder of the paper, we choose a large tree size rather than use cross-validation, since this seems to produce the most competitive RP models.

To understand the two results, we first review cross-validation and cost-complexity pruning. Breiman et al.[9] presented a cross-validated cost-complexity pruning approach to selecting the best tree size. First, a large tree is grown using the training data. A nested sequence of m pruned trees (with $s_1 < s_2 < ... < s_m$ terminal nodes) is generated, minimizing a cost-complexity criterion. The goal is then to choose one tree from this nested sequence (that is, choose $s^* \in s_1,...s_m$). This is accomplished by cross-validation. For example, with 10-fold cross-validation, 10 different large trees are grown and pruned, yielding 10 nested sequences, with sizes corresponding to $s_1,...s_m$. Each of the 10 different trees are grown using 90% of the training data and holding out a different validation set of 10% of the training data. For each tree size s_i, prediction errors are averaged across the 10 validation folds, and the best tree size s^* is chosen so as to minimize this cross-validation error.

For the AIDS assay data and 64 BCUT descriptors, we generated 20 training sets of 15 000 observations, about half of the data set, and for each set chose the optimal tree size according to cross-validated error. In order to see whether we are choosing the right size, we "cheat" by using the remaining (approximately 15 000) observations as a test set to choose the right tree size. The tree sizes chosen by these two strategies are displayed in Figure 3. Cross-validation typically selects a tree with between 1 and 100 terminal nodes, occasionally selecting a tree with around 150 terminal nodes. In contrast, the test set reveals that the best tree size is never below 60 terminal nodes and is usually greater than 150 terminal nodes. The choice by cross-validation of a too-small tree results in a decrease in prediction accuracy for the test set, in comparison to the optimal tree size. Note that the use of a test set to choose tree size is generally inappropriate and is used here simply to illustrate that cross-validation seems to select an inappropriate sized tree.

A possible reason for this discrepancy is the large number of "ties" that a tree produces in its predictions for the probability of activity. All observations falling in a specific terminal node of the tree will receive the same prediction. Tie structure among the predictions can affect predictive accuracy in unbalanced response problems where ranking is the goal. In cross-validation, since only 10% of the data are predicted by a tree corresponding to each fold, and the trees are slightly different for each fold, there are fewer ties in the predictions. This may make a small tree appear to be a better performer under cross-validation, where it seems to generate fewer ties, than for an independent test set, where one tree is generated from all the training data, with more ties in predictions.

3. EVALUATION PLAN FOR CSARA AND RP

In section 3.1 we describe three strategies for data sampling/analysis based on CSARA, RP, and a hybrid method and the performance metric for their evaluation. The two assays and the various descriptor sets used for empirical comparisons are then outlined in section 3.2.

3.1. Three Sampling/Analysis Strategies and Their Evaluation. All three strategies to be described will be evaluated in terms of their performance in identifying active compounds in the second round of HTS.

Specifically, we will use the hit rate (HR), which is the proportion of compounds that are assayed to be biologically active among a specified group of compounds predicted to be active. As CSARA is a two-stage process of sequential screening, hits (active compounds) and hence the hit rate come from both the first and the second screens. A tree generated by RP, however, is used to make predictions and hence choose compounds for only the second screen. Thus, all comparisons of HR will be made only for the second screen. The assay results from the first screen are treated as "training" data, and those from the second screen are treated as "test" data. Nonetheless, the performance of RP will depend on the training data used to fit it. Thus, we will also consider the role of training data in defining strategies.

Two distinct ways are used to generate training data:

The Cluster method is inherent to CSARA, but it can also be used to generate training data for RP. It allows us to evaluate the differences between CSARA and RP when the same training data are used in modeling the QSAR. The Random method can be used only for RP. It will allow us

Methods for Sampling Training Data

Cluster: The compound collection is clustered into K clusters, and one compound is randomly chosen from each cluster. This is Step B of the CSARA algorithm, illustrated in Figure 2 (B).

Random: K compounds are chosen completely at random from the collection.

to assess the usefulness of a training sample designed to be diverse versus a random selection.

Hence, there are three different data sampling/analysis strategies:

Three Sampling/Analysis Strategies

CSARA: All steps of the CSARA method as described in Section 2.1.

Cluster/RP: The Cluster sampling method is used to select the initial sample set, an RP tree is trained on this set, and predictions from the tree model are used to choose the follow-up sample.

Random/RP: The Random sampling method is used to select the initial sample set, an RP tree is trained on this set, and predictions from the tree model are used to choose the follow-up sample.

The comparisons of interest are CSARA versus Cluster/RP and Random/RP, to understand the impact of the QSAR modeling method, and Cluster/RP versus Random/RP to understand the impact of the sampling of the training set.

3.2. Data Sets. Two assays are used in this paper with a variety of descriptor sets, as summarized in Table 1.

The first set of assay data is from the National Cancer Institute (NCI) Yeast Anticancer Drug Screen[11] and will be referred to here as the yeast data. The yeast assay measures inhibition of tumor growth. For a tumor to develop, there must be a series of mutations, which cause cells to multiply uncontrollably. Because of the high degree of functional homology in biological activity between yeast and mammalian cells, many mutations can be modeled in yeast. Simon et al.[11] carried out a screen of over 100 000 compounds from the repository at the NCI's Developmental Therapeutics Program to identify compounds that can inhibit the growth of the mutated cells. Hence, the yeast assay data measure the percentage growth inhibition of the assayed compounds. Among 100 000 compounds screened, 75 873 are used in the analysis. For most analyses we will convert the percentage inhibition to a binary inactive/active response. Of the 75 873 compounds, 6834 are considered active, because they have growth inhibition of at least 70%;[11] the proportion of active compounds is 8.2%. The yeast data can be downloaded from http://dtp.nci.gov/yacds.[15]

The second assay, from the NCI Developmental Therapeutics Program, relates to the HIV/AIDS virus and will be called the AIDS data. A description of this assay is provided by Lam et al.[12] The original AIDS data can be downloaded from http://dtp.nci.nih.gov/docs/aids/aids_data.html.[14] Some observations with poor structure representations that are usually considered as nondrug candidates have been deleted from the original data.[12] The biological activity is the amount of protection a compound gives to human CEM cells from HIV-1 infection. Two assay classifications, moderately active and confirmed active, have been combined to form an "active" class.[13] There are approximately 2% actives. The data set and its structure file are available from the authors by request.

The data for both assays were generated by HTS. When converted to a binary outcome, the yeast data have a higher proportion of active compounds than found in the AIDS data. In general, these compounds are representative of those in pharmaceutical data sets. Although they are not completely typical as they include "toxic" compounds, which may selectively kill cancer cells, they should still provide useful information on the effectiveness of the QSAR sampling/analysis strategies.

In all experiments, we will treat the available data as if they were a compound library and pretend to observe activity for selected subsets of the library.

Various descriptor sets, as summarized in Table 1, will be used to characterize chemical structure in the QSAR modeling.

4. EXPERIMENT AND RESULTS

4.1. Yeast Data. First, we apply CSARA, Cluster/RP, and Random/RP to the yeast data. We need to specify K, the number of compounds from the first screen to be used for training data: We take $K = 3000$, 7000, and 15 000. Because all methods considered rely on some form of randomization to select the training data, four trials are run at each of the three levels of K. Six BCUT descriptors are used.

The results are presented in Tables 2–4. As the results follow similar patterns across values of K and the four trials, we explain in detail only the first row of Table 2. The first column is the trial index. The second and third columns correspond to the first screen using steps A–C of the CSARA method. After the first screen, there are 243 hits in the training set, giving a hit rate of 243/3000 = 8.1%. The fourth to sixth columns give results from the second HTS using step D of CSARA: All compounds in the active clusters (not including those in the first screen) are treated as the test data for the second round HTS. In this example, 5648 compounds are in the test data, there are 874 hits, and consequently the hit rate is 874/5648 = 15.5%.

The seventh and eighth columns give results from the second screen using Cluster/RP. In the first row of Table 2, Cluster/RP selects the 5648 compounds with the highest predicted probabilities of being active (the size of the test set is matched to CSARA's). Roughly 819 or 14.5% are active. The same technique of selecting for the second screen is used for Random/RP in columns 9 and 10, but the training data are $K = 3000$ randomly chosen compounds.

For tree models, calculation of the number of hits is complicated by the presence of many tied predictions. All test points falling in the same terminal node will receive the same predicted probability of activity. For example, suppose the best 14 nodes give us 5448 compounds from the test set, and the 15th best node has an additional 300 test set compounds. We need to select 200 of these 300 compounds to give the desired 5648 compounds. We deal with this problem by reporting an expected number of hits under random sampling of 200 compounds from the 300 available in the node. This is equivalent to linear interpolation of number of hits between the two nodes.

In Tables 2–4, within each row, CSARA and Cluster/RP can be compared since they have the same training data set. Comparing Random/RP with CSARA or Cluster/RP is not meaningful within a row because they have different training

Table 1. Data Sets Used in the Paper

assay			descriptor variables		
name	measurement	no. of compds	set	no.	source
yeast	binary	75 873	BCUT	6	GlaxoSmithKline
	continuous	75 873	BCUT	6	GlaxoSmithKline
AIDS	binary	29 812	BCUT	6	GlaxoSmithKline
	binary	29 374	BCUT	64	Feng et al.[16]
	binary	29 374	constitutional (CON)	46	Feng et al.[16]
	binary	29 374	property (PROP)	212	Feng et al.[16]
	binary	29 374	topological (TOP)	261	Feng et al.[16]

Table 2. Results for the Yeast Data and $K = 3000$ Clusters

	first screen		second screen (CSARA)			Cluster/RP		Random/RP	
trial	hits	HR (%)	no. of compds	hits	HR (%)	hits	HR (%)	hits	HR (%)
1	243	8.1	5648	874	15.5	819	14.5	676	12.0
2	250	8.3	5720	861	15.1	870	15.2	753	13.2
3	261	8.7	6060	841	13.9	678	11.2	788	13.0
4	262	8.7	6272	1020	16.3	917	14.6	777	12.4
av	254	8.47	5925	899	15.2	821	13.9	749	12.6

Table 3. Results for the Yeast Data and $K = 7000$ Clusters

	first screen		second screen (CSARA)			Cluster/RP		Random/RP	
trial	hits	HR (%)	no. of compds	hits	HR (%)	hits	HR (%)	hits	HR (%)
1	584	8.3	5783	1056	18.3	888	15.4	960	16.6
2	584	8.3	5785	1069	18.5	938	16.2	857	14.8
3	580	8.3	5515	1023	18.6	872	15.8	809	14.7
4	583	8.3	5735	1053	18.4	890	18.4	877	15.3
av	583	8.3	5705	1050	18.4	897	16.4	876	15.3

Table 4. Results for the Yeast Data and $K = 15\,000$ Clusters

	first screen		second screen (CSARA)			Cluster/RP		Random/RP	
trial	hits	HR (%)	no. of compds	hits	HR (%)	hits	HR (%)	hits	HR (%)
1	1182	7.9	4771	1066	22.3	898	18.8	845	17.7
2	1321	8.8	5417	1110	20.5	928	17.1	955	17.6
3	1259	8.4	5197	1064	20.5	871	16.8	894	17.2
4	1286	8.6	5099	1057	20.7	871	17.1	860	16.9
av	1262	8.4	5121	1074	21.0	892	17.5	889	17.4

sets. Comparisons using the "average" ("av" denoted in Tables 2−4) row are valid between all three methods.

The results given in Tables 2−4 can be summarized as follows:

- As K increases, the hit rate in the test set also increases. This makes intuitive sense because a larger training set gives more information to the tree and clustering algorithm, enabling them to uncover the QSAR within the HTS data. Also, as K increases the change in the hit rate is larger for CSARA than for Random/RP and much larger than for Cluster/RP.
- Whatever K is chosen, CSARA always has a higher hit rate than Cluster/RP and Random/RP. This indicates that CSARA outperforms Cluster/RP and Random/RP.

4.2. AIDS Data. A similar analysis is carried out for the AIDS data, but $K = 3000$, 5000, 7000, 10 000 and 15 000, and 20 trials (training sets) are used. Again, the descriptor set is six BCUTs.

Table 5. Results for the AIDS Data

	CSARA		Cluster/RP		Random/RP	
K	mean HR (%)	(SE) (%)	mean HR (%)	(SE) (%)	mean HR (%)	(SE) (%)
3000	20.3	(0.6)	15.6	(0.7)	15.6	(0.8)
5000	24.5	(0.7)	19.5	(0.8)	18.0	(0.6)
7000	26.4	(0.4)	21.3	(0.8)	20.6	(0.5)
10 000	31.6	(0.5)	26.7	(0.7)	22.5	(0.5)
15 000	36.6	(0.5)	30.0	(0.5)	24.7	(0.5)

Instead of presenting results for each trial, as in Tables 2−4, we report in Table 5 averages and standard errors across the 20 sets. Since each row represents an average over multiple training sets, all three methods can be compared within a row.

Table 5 shows that

- The standard errors of the hit rate are very small, implying that 20 trials are sufficient to compare the mean hit rates of the three sampling/analysis strategies.
- For all values of K considered, the mean hit rate of CSARA is consistently larger than that of Cluster/RP or Random/RP. Thus, the CSARA method is competitive and efficient relative to RP for identifying active compounds here.
- Cluster/RP outperforms Random/RP, particularly for larger values of K: A diverse training set is advantageous here.

Formal statistical tests, i.e., paired *t*-tests for CSARA versus Cluster/RP and unpaired *t*-tests for the other comparisons, confirm the above findings at a 5% significance level. The differences in mean HR are statistically significant for CSARA against either RP competitor for all values of K and for Cluster/RP against Random/RP for $K = 10\,000$ and 15 000.

4.3. Adding Irrelevant Descriptors. In order to test the stability of CSARA, we add several irrelevant or "junk" descriptor variables to the six BCUT descriptors when modeling the AIDS data. The values of the first new, irrelevant descriptor are generated by randomly permuting the values of the first BCUT, the second irrelevant variable is generated in the same way from the second BCUT, and so on. We take $K = 15\,000$ in this experiment.

Table 6 illustrates the effect of adding 1−6 junk variables. The hit rates reported are means over four trials. With more irrelevant descriptors, the mean hit rate of CSARA decreases much more quickly than that of Cluster/RP or Random/RP. Among these three methods, Random/RP is the most stable. RP has built-in variable selection due to choosing a variable at each split according to an optimality criterion (here the deviance). CSARA has no such capability, as all descriptors are included in the distance metric for clustering. Similarly,

Table 6. Mean Hit Rate for the AIDS Data and $K = 15\,000$ When Irrelevant Descriptors Are Added to the Six BCUTS

no. of irrelevant descriptors	CSARA mean HR (%)	Cluster/RP mean HR (%)	Random/RP mean HR (%)
0	36.6	30.0	24.7
1	26.2	28.7	24.1
2	20.5	26.8	23.8
3	14.0	23.7	22.8
4	12.1	23.2	25.7
5	9.5	23.6	21.3
6	8.2	19.5	20.5

the benefit due to clustering in selection of a training set for RP diminishes with more irrelevant variables.

5. EXPERIMENTS WITH HIGH-DIMENSIONAL DESCRIPTOR SETS

In order to understand how CSARA performs with higher-dimensional descriptor sets, we make comparisons between CSARA, Cluster/RP, and Random/RP using the AIDS assay data and four further descriptor sets. The further sets, summarized in Table 1, are BCUT (64 variables), constitutional (46 variables), property (212 variables), and topological (261 variables). More details are available at http://stat.ubc.ca/~will/ddd/.[17] All these descriptors were computed by Feng et al.[16] using an old version of DRAGON, which is no longer available. The current version of DRAGON is available at http://www.talete.mi.it/dragon_exp.htm.[18]

BCUT descriptors are determined by the connectivity, i.e., the 2-D topological relations between the heavy atoms of a molecule. An adjacency matrix, also called the Burden matrix, is used to calculate BCUTs. In this matrix, atomic properties are placed on the major diagonal and a measure of connection is placed in the off-diagonal cells. Feng et al.[16] considered four atomic properties—atomic mass, van der Waals volume, atomic electronegativity, and atomic polarizability. The eight largest and the eight smallest eigenvalues of each atomic property are used, so there are in total $16 \times 4 = 64$ descriptors.

Constitutional descriptors are independent of molecular connectivity and conformations of the molecule. These descriptors are determined by molecular properties such as atom and bond counts, molecular weight, atomic number, etc. Property descriptors reflect physicochemical properties of molecules, like log P, aromatic index, etc. They also include fragment descriptors, which indicate the kinds of fragments in a molecule and their frequencies. The fragments include atom/bond sequences and augmented atoms. Topological descriptors are obtained from the molecular graph and do not depend on conformation. Topological descriptors are easy to calculate, but sensitive to small changes in molecular structure.

As before, we run 20 trials for different values of K and calculate the mean hit rates with standard errors.

Figure 4 shows that CSARA outperforms Cluster/RP and Random/RP. These high-dimensional results may seem to conflict with the experiment in section 4.3, where CSARA performance degraded quickly with the addition of further, irrelevant descriptors. CSARA's strong performance here with high-dimensional sets is probably a reflection of the quality of the sets, where all variables may be at least weakly informative. In contrast, irrelevant variables are completely unrelated to activity.

Table 7 displays the significance levels for tests of a difference in mean HR, comparing CSARA, Cluster/RP, and Random/RP pairwise for each of the four descriptor sets. It is clear that CSARA has a significantly larger mean HR than Cluster/RP or Random/RP and that Cluster/RP performs better than Random/RP as K increases.

6. APPLICATION OF CSARA TO A CONTINUOUS ASSAY RESPONSE

As mentioned in section 3.2, growth inhibition (potency) of compounds was originally measured as a continuous response for the yeast assay. “Active” and “inactive” labels were obtained by thresholding the response. Here, we adapt CSARA for a continuous response and compare performance with RP methods.

Step D of the CSARA algorithm in section 2.1 is adapted as follows. Each cluster is scored according to the potency of the compound randomly sampled from it for assay. All compounds in the highest scoring cluster are chosen first for the second-round assay, then those in the second-highest scoring cluster, and so on. Building an RP regression tree for a continuous response is well-known.[9] Comparisons are carried out with the six BCUT descriptors and $K = 3000$.

The results are shown graphically in Figure 5. The average potency of the selected compounds is plotted against the number of compounds selected. A curve that is high at the left and decreases gradually would indicate good ability to identify high potency compounds.

Several observations can be made on the basis of Figure 5:

- Up to about 1000 compounds selected, there is substantial improvement over random selection, which would correspond to a horizontal line at a height of approximately 8%.
- Compared with the RP methods, performance of CSARA falls off faster with the number of compounds selected.
- If 70% predicted inhibition is used as a cutoff, CSARA chooses more compounds for the second screen compared with the RP strategies.

Similar results (not shown) are obtained for $K = 10\,000$, except that all strategies, not surprisingly, return higher average potencies for the first 10 000 compounds selected with a larger K.

7. DISCUSSION AND FUTURE RESEARCH

The main aim of this paper is exploring the properties of CSARA. Two data sets are analyzed to help evaluate the differences between CSARA and RP. The results suggest that CSARA outperforms RP models in selecting more hits if the assay is a binary outcome (active/inactive). RP trees are trained to give overall good prediction, giving equal weight to both active and inactive compounds. In contrast, in its very simple analysis of the first-round training data, CSARA gives heavy weight (100%) to active compounds and no weight to inactive compounds. Thus, CSARA may be a more appropriate method for drug discovery data sets where actives are rare.

However, the largest limitation with CSARA is its instability. When there are many irrelevant descriptors, the effectiveness of CSARA decreases. In the presence of many

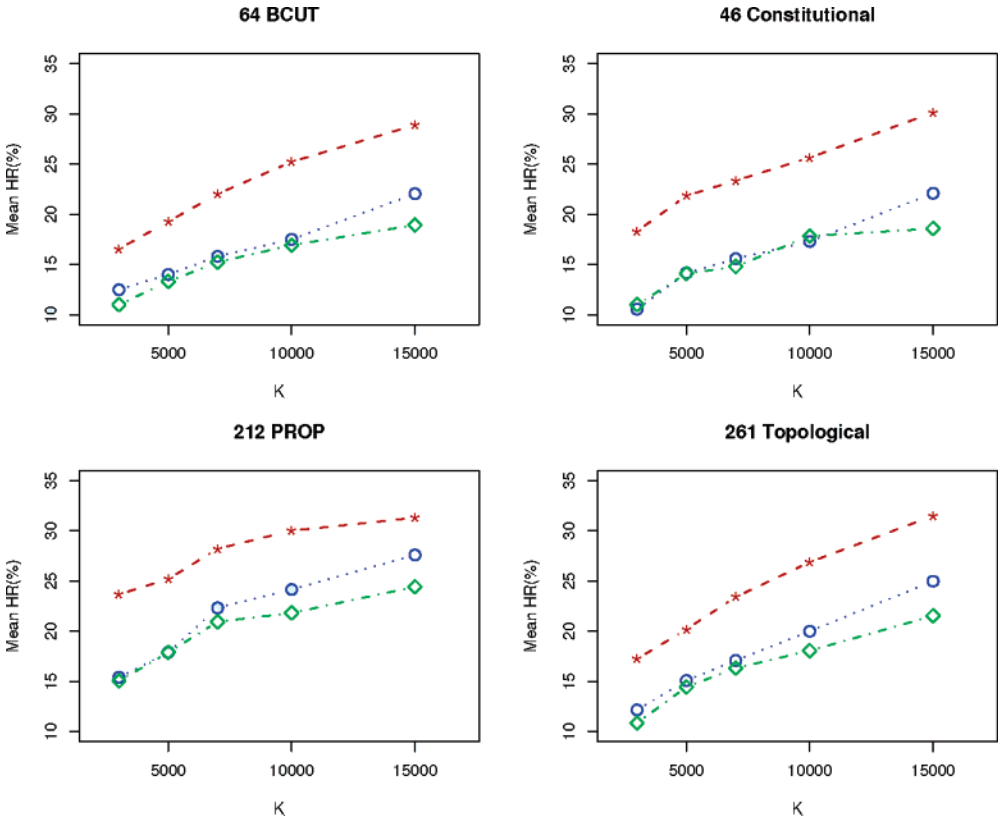

Figure 4. Mean hit rate versus *K* for CSARA (★···★), Cluster/RP (○···○), and Random/RP (◇···◇) for the AIDS data with four high-dimensional descriptor sets.

Table 7. Hypothesis Tests Comparing Mean HR for Three Sampling/Analysis Strategies When Applied to the AIDS Data with Four Descriptor Sets: 64 BCUT Descriptors, 46 Constitutional Descriptors, 212 Property Descriptors, and 261 Topological Descriptors[a]

	CSARA vs Cluster/RP				CSARA vs Random/RP				Cluster/RP vs Random/RP			
K	BCUT	CON	PROP	TOP	BCUT	CON	PROP	TOP	BCUT	CON	PROP	TOP
3000	***	***	***	***	***	***	***	***				
5000	***	***	***	***	***	***	***	***				
7000	***	***	***	***	***	***	***	***				
10 000	***	***	***	***	***	***	***	***			**	*
15 000	***	***	***	***	***	***	***	***	***	***	***	***

[a] Differences in mean HR significant at the 5%, 1%, and 0.1% levels are denoted by *, **, and ***, respectively.

potential irrelevant descriptors, variable selection may need to be carried out first. Here the term "irrelevant" variable refers to those that have no relationship with the biological assay and hence have a negative impact on QSAR model training and prediction. For the AIDS assay data and the four high-dimensional descriptor sets used in the paper, some experiments have been done to test if there is any performance improvement of CSARA, Cluster/RP, and Random/RP using the important variables identified by Partitionator from www.goldenhelix.com. No significant improvements were found for the three approaches. It is not surprising that the performance of Cluster/RP and Random/RP is not improved by variable selection, as trees can automatically select important variables at each split. Here we need to emphasize that the variables that are not chosen as important variables are not necessarily irrelevant. They seemingly do not help in the prediction of biological activity, but they are not harmful to prediction either, possibly because of correlations with other variables. This may explain why variable selection does not improve CSARA for these descriptor sets.

One potential shortcoming of CSARA is the lack of control over the number of compounds selected for a second HTS. In applications where the categorical activity is formed from an underlying continuous response, the methods in section 6 may circumvent this difficulty.

The results in section 6 also provide much insight into the comparisons between CSARA and the RP methods. RP trees may be effective in choosing a relatively small number of second-screen (test) compounds. In most of the experiments in this paper, however, we matched the number of test compounds across CSARA and the RP methods, forcing RP to select a large number of compounds if CSARA does. When control over the size of the second screen is made possible for CSARA too (by working with a continuous assay measurement), RP methods can become more competitive for smaller screens.

In this paper, exactly one compound is randomly selected from each cluster, regardless of cluster size. In future research, we will consider sampling more than one compound

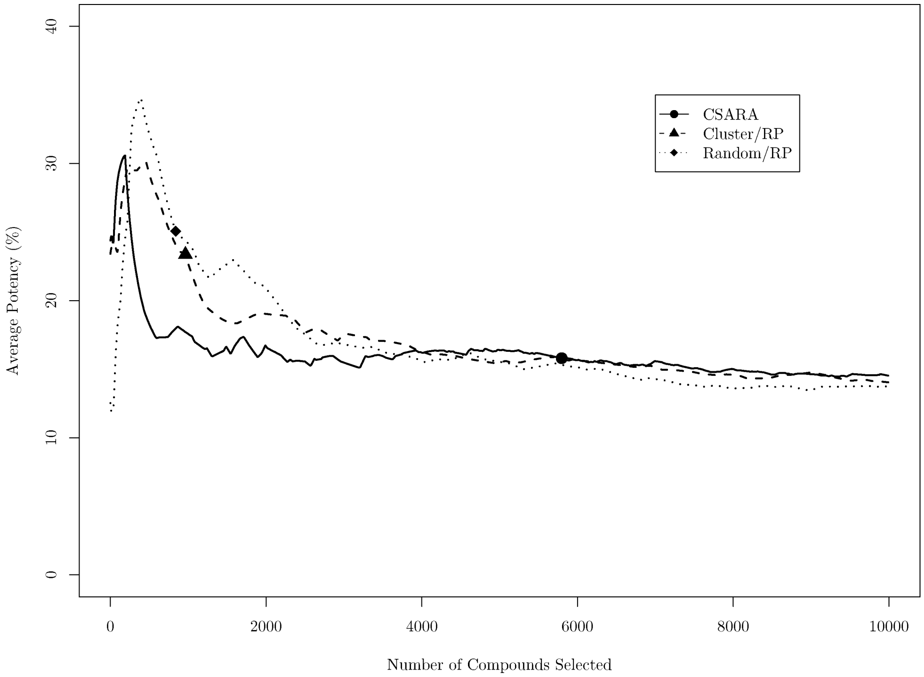

Figure 5. Average potency (%) of selected compounds versus the number of compounds selected for CSARA, Cluster/RP, and Random/RP, when $K = 3000$. The symbols on the curves show where selection would stop if a predicted potency larger than 70% inhibition is required.

per cluster and allowing the number of compounds sampled to vary according to cluster size.

ACKNOWLEDGMENT

This research was supported by Mathematics of Information Technology and Complex Systems (Canada), the NIH Roadmap for Medical Research, and the Natural Sciences and Engineering Research Council (Canada). We thank ISC BioExpress for permission to use a photograph of one of their products as the background of our table of contents.

APPENDIX: CHOICE OF THE BEST TREE SIZE

This Appendix outlines some technical details in the experiments on tree size selection.

The performance measure used to assess tree performance is the hit rate or the percentage of hits among those compounds selected. In order to facilitate comparisons with CSARA, the number of compounds selected is matched in each experiment with the number selected by CSARA.

Ideally, a tree would be grown and pruned according to the hit rate. However, our goal is to compare “off-the-shelf” versions of tree growing algorithms with CSARA, and the hit rate is not a standard criterion for growing or pruning a tree. Thus we choose the Gini index, $\sum_{i=1}^{n} \hat{p}_i(1 - \hat{p}_i)$, where $\hat{p}_i$ is the predicted probability of activity for observation i, as a surrogate measure for tree growing and pruning. The Gini index is among the most appropriate measures since it encourages models to accurately predict p_i, the probability of activity for observation i, rather than just predict a class label. Once a sequence of pruned trees has been generated using the surrogate pruning criterion, predictions can be generated and the hit rate (the real performance measure) can be evaluated. Thus a cross-validated measure of the hit rate will be obtained for each tree in the nested sequence of trees. We explore two different ways of generating this hit rate: either calculating a hit rate separately for each of the ten folds, and averaging them, or combining the predictions from all ten validation sets and then ranking the compounds and generating a single hit rate. Similar results were obtained using either strategy.

All RP calculations were carried out in R,[19] using the rpart library.[20]

REFERENCES AND NOTES

(1) Drews, J. Drug Discovery: A Historical Perspective. *Science* **2000**, *287*, 1960−1965.

(2) Valler, M. J.; Green, D. Diversity screening versus focussed screening in drug discovery. *Drug Discovery Today* **2000**, *5*, 286−293.

(3) Abt, M.; Lim, Y.-B.; Sacks, J.; Xie, M.; Young, S. S. A Sequential Approach for Identifying Lead Compounds in Large Chemical Databases. *Stat. Sci.* **2001**, *16*, 154−168.

(4) Engels, M. F. M.; Venkatarangan, P. Smart Screening: Approaches to Efficient HTS. *Curr. Opin. Drug Discovery Dev.* **2001**, *4*, 275−283.

(5) Lajiness, M. S. Dissimilarity-based compound Selection Techniques. *Perspect. Drug. Discovery Des.* **1997**, *7*, 65−84.

(6) Dunbar, J. B. Cluster-based Selection. *Perspect. Drug Discovery Des.* **1997**, *7/8*, 51−63.

(7) MacQueen, J. *Some Methods for Classification and Analysis of Multivariate Observations*, Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, Le Cam, M. L., Neyman, J., Eds.; University of California Press: Berkeley, CA, 1967; Vol. 1.

(8) Hartigan, J. A.; Wong, M. A. Algorithm AS136: A K-means Clustering Algorithm. *Appl. Stat.* **1979**, *28*, 100−108.

(9) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. Chapter 3: Right Sized Trees and Honest Estimates (59-81). *Classification and regression trees;* Wadsworth Publishing Co. Inc.: 1984.

(10) Rusinko, A.; Farmen, M. W.; Lambert, C. G.; Brown, P. L.; Young, S. S. Analysis of a Large Structure/Biological Activity Data Set Using Recursive Partitioning. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1017−1026.

(11) Simon, J. A.; Dunstan, H.; Lamb, J. R.; Evans, D. R.; Cronk, M.; Irvine, W. 015 Yeast as a model organism for anticancer drug discovery: An update from the NCI/Fred Hutchinson Cancer Research Center collaboration, Proceedings of the 11th NCI · EORTC · AACR Symposium; 2000.

(12) Lam, R. L. H; Welch, W. J.; Young, S. S. Uniform Coverage Designs for Molecule Selection. *Technometrics* **2002**, *44*, 99−109.

(13) Lam, R. L. H. Design and Analysis of Large Chemical Databases for Drug Discovery, Thesis, Department of Statistics and Actuarial Science, University of Waterloo, Canada, 2001.

(14) NCI AIDS Antiviral Drug Screen data. http://dtp.nci.nih.gov/docs/aids/aids_data.html (accessed April 26, 2003).

(15) NCI Yeast Anticancer Drug Screen data. http://dtp.nci.nih.gov/yacds (accessed April 26, 2003).

(16) Feng, J.; Lurati, L.; Ouyang, H.; Robinson, T.; Wang, Y. Y.; Yuan, S.; Young, S. S. Predictive Toxicology: Benchmarking Molecular Descriptors and Statistical Methods. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1463−1470.

(17) Chipman, H. A.; Shen, H.; Welch, W. J.; Wang, M.; Young, S. S.; Yuan, F. Drug Discovery Data Benchmarking Resource Centre. http://stat.ubc.ca/~will/ddd/ (accessed Jan 6, 2006).

(18) Talete SRL. DRAGON software. http://www.talete.mi.it/dragon_exp.htm (accessed Dec 28, 2006).

(19) R Development Core Team. *R: A language and environment for statistical computing*; R Foundation for Statistical Computing: Vienna, Austria, 2006; ISBN 3-900051-00-3.

(20) Therneau, T. M.; Atkinson, B. *rpart: Recursive Partitioning, 2006 R package version 3.1-29*; R port by Brian Ripley.

CI600458N