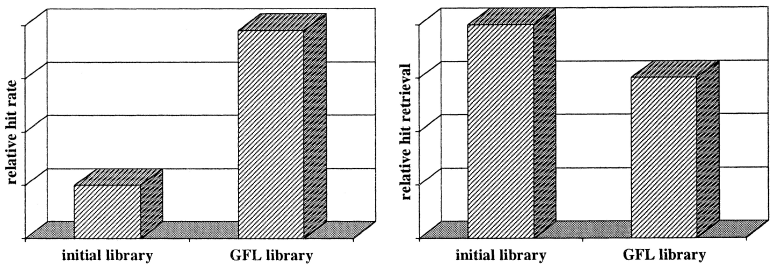
JOURNAL OF combinatorial CHEMISTRY

Article

Retrospective Analysis of an Experimental High-Throughput Screening Data Set by Recursive Partitioning

A. Michiel van Rhee, Jon Stocker, David Printzenhoff, Chris Creech, P. Kay Wagoner, and Kerry L. Spear

J. Comb. Chem., **2001**, 3 (3), 267-277• DOI: 10.1021/cc0000747 • Publication Date (Web): 28 February 2001



More About This Article

Additional resources and features associated with this article are available within the HTML version:

- Supporting Information
- Links to the 1 articles that cite this article, as of the time of this article download
- Access to high resolution figures
- Links to articles and content related to this article
- Copyright permission to reproduce figures and/or text from this article

View the Full Text HTML

Journal of Combinatorial Chemistry is published by the American Chemical Society.
1155 Sixteenth Street N.W., Washington, DC 20036

Retrospective Analysis of an Experimental High-Throughput Screening Data Set by Recursive Partitioning

A. Michiel van Rhee,* Jon Stocker, David Printzenhoff, Chris Creech, P. Kay Wagoner, and Kerry L. Spear

ICAgen, Inc., P.O. Box 14487, Research Triangle Park, North Carolina 27709

Received August 25, 2000

With the emergence of combinatorial chemistry, whether based on parallel, mixture, solution, or solid phase chemistry, it is now possible to generate large numbers of diverse or focused compound libraries. In this paper we aim to demonstrate that it is possible to design targeted libraries by applying nonparametric statistical methods, recursive partitioning in particular, to large data sets containing thousands of compounds and their associated biological data. Moreover, when applied to an experimental high-throughput screening (HTS) data set, our data strongly suggest that this method can improve the hit rate of our primary screens (about 4- to 5-fold) while increasing screening efficiency: less than one-fifth of the complete selection needs to be screened in order to identify about 75% of all actives present.

Introduction

In recent years, combinatorial chemistry coupled with high-throughput screening (HTS) has dramatically increased the number of compounds that are screened against many biological targets. Despite the resulting explosion of screening data for a given target, hit rates still tend to be quite low (often less than 1%). Our interest in discovering novel, small molecule modulators (blockers, openers, or otherwise) of ion channels has directed our attention to exploring methods for improving hit rates beyond those obtained with historically, randomly or diversely chosen compound collections.

Ion channels are membrane embedded proteins of multimeric composition with intrinsic ion conduction properties. The intended pharmacological endpoint, i.e., activation, prolongation of activation, termination of activation, or block of the target ion channel, is dependent on the site and mode of binding of the ligand to the channel. Consequently, we are forced to reevaluate the leading pharmacological paradigm that we can competitively displace a natural or xenobiotic agent from a binding site. Therefore, to further our understanding of ion channel biology and medicine, we wanted to find an efficient way to identify modulators of multiple ion channel subtypes within ion channel families, so-called gene families, without focusing on a single binding site or mechanism.

Our approach needed to encompass the following: (1) the ability to increase the efficiency of our primary screens, (2) the option to pursue multiple chemotypes in order to develop compounds along parallel product lines, and (3) the ability to explain nonlinear structure−activity relationships. Various deterministic methods have been applied to try to resolve similar problems.[1−8] The most limiting condition for these methods, however, is probably the requirement of the training set to encompass all chemotypes present in the test set, the so-called knowledge domain. Whereas this may not be of concern when discerning drug-like from nondrug-like compounds based on sufficiently diverse databases such as ACD,[9] CMC,[10] or WDI,[11] this becomes a self-limiting "conservatism in design" restriction[12] when screening for as yet unidentified activity in a narrowly defined chemical library.

After review and testing of several possible solutions, such as ANOVA, hierarchical cluster analysis, principal components analysis, factor analysis, genetic function approximations, partial least squares fitting, multiple pharmacophore based models, and combinations thereof, we explored nonparametric methods, and recursive partitioning in particular. Methodologically, we can distinguish parametric methods[13−17] (i.e., a combination of chemical descriptors explains and predicts the biological activity of each compound in the training set) from nonparametric methods (i.e., one can calculate the chance of a compound being within a range of biological activities based on the distribution of chemical and biological descriptors in the training set). The limitation of all these QSAR methods is that a single (quasi-) linear equation is presumed to account for all biological activity. Whereas this may hold true for selective, reversible, and competitive binding models, these conditions need not necessarily apply to HTS data sets. Furthermore, past research here and elsewhere[18−20] indicates that it is very likely that chemical modulators of ion channels, especially those that are endogenously regulated by membrane potentials (e.g., the K_v gene family) or ion concentrations (e.g., Ca^{2+} and Cl^- channels), are noncompetitive, or uncompetitive, allosteric modulators. Therefore, it becomes imperative that analysis methods are applied that allow for the presence and/or selection of multiple binding mode models, rather than converge on a single unified model.

* To whom correspondence should be addressed.

10.1021/cc0000747 CCC: $20.00
Published on Web 02/28/2001

Recursive partitioning (RP) is a nonparametric classification technique that has been shown to have applicability in, e.g., a clinical setting.[21−25] In studies designed to identify risk subgroups, RP successfully identified subgroups with distinct risk assessments that had previously not been identified using more traditional logistic regression.[21,26] This finding agrees with our assertion that RP may be able to identify discreet binding modes within an HTS data set. An overview and comparison of various statistical methods, e.g., linear discriminant analysis, logistic regression, nearest neighbor clustering, and recursive partitioning, were presented by Hand several years ago.[23]

With the advent of combinatorial chemistry and HTS, the data structure and organization increasingly seem to more closely resemble a rather disparate patient population than an idealized lead optimization set such as envisioned by Topliss[27,28] and employed by many medicinal chemists. Comparisons of nonparametric recursive partitioning to parametric analyses have been performed[22,25] and generally indicate that RP is significantly better at identifying synergistic and nonlinear relationships, whereas multivariate techniques perform better at late stage analyses with more homogeneous data sets.

RP is a method whereby a group of compounds is recursively (i.e., starting with the complete set and ending with the smallest possible or allowable subset) split at a branch point into two statistically distinct nodes (subsets).[29] Whereas variable selection in parametric methods is determined by their impact on correlation, RP focuses on classification. As such, RP has the possibility to optimize for synergism rather than additivity, for nonlinear relationships over forced (quasi-) linearity, and for multiple endpoints over single endpoints. In addition, during variable selection RP takes into account the prior probabilities and penalties for misclassification. In contrast, RP has diminishing numbers in each discriminant step, whereas parametric methods retain all information elements during the equation building phase. The most significant drawback to the application of RP is perhaps that it may underestimate the predictive ability of linear and continuous factors.[22]

Recursive partitioning itself was demonstrated by Young and Hawkins,[30] as early as 1995, to be a powerful method of harnessing the information content of a combinatorial chemical library. The size of the chemical library, a 2^9 factorial design of 512 compounds, and a well-defined target set the precedent for the entire field. Whereas a typical SAR series usually comprises as few as 20 or as many as 50 compounds, this approach increases the dimension of the problem at least by an order of magnitude. At the conclusion of their paper, the authors mused "If there were more chemical components at each position and the components were described with many numerical descriptors, then the analysis problem would be more difficult (and realistic). The problem would be much more difficult (impossible?) if the set of compounds was some sort of catch-as-catch-can collection."

Rusinko et al.[31] recently reported on a particular implementation of recursive partitioning, i.e., SCAM (statistical classification of the activities of molecules), to identify structure−activity relationships in large data sets. They documented how they were able to effectively use this method to develop a structure−activity relationship (SAR) for a set of 1650 compounds with 6405 binary descriptors and obtain up to a 15-fold enhancement over the random hit rate while virtually "screening" the WDI.[9]

Our study differs from these earlier reports, and those using other data sets,[8,9,32] in that we use both combinatorial chemistry and experimental data derived from an HTS assay. Furthermore, we present evidence that we can increase the scale of the problem by another 1 to 2 orders of magnitude and obtain significant results. In this paper, therefore, we aim to prove that RP is an effective method in harnessing the chemical and biological information contained in a chemical library and its HTS data, even in the face of the odds presented above.

Methods

A 20 986-member library from our compound inventory was selected and submitted for screening. This chemical library was entirely composed of combinatorial chemistry derived compounds, synthesized either by solid or by liquid phase parallel methods. The biological activity of all library members was determined individually. A set of eight randomly selected plates, accounting for 640 library members, was analyzed by LC/MS. Of the total number of samples analyzed, 81% were found to be better than 80% pure, and 66% were found to be better than 90% pure. The median, average, and standard deviation values were 94%, 88%, and 15%, respectively. Therefore, the purity of the majority of the library members was deemed to exceed 80%. The combinatorial process was directed by synthetic feasibility without prior knowledge of the biological target. Since the chemical library was set up to take advantage of synthetic feasibility rather than molecular diversity, no diversity analysis prior to compound selection was performed.

The set was mathematically divided into a 5000-member training set based on either diverse selection (DS; D-optimal design strategy) or random selection (RS; iteratively obtained using a random number generator) and into a 15 986-member test set. Biological data were generated in a high-throughput screening (HTS) fashion using a cell-based method proprietary to ICAgen, Inc. The library members were subsequently assigned to activity bins based on their relative biological activity. For the quaternary analysis: 147 in class 4 ("highly active"), 471 in class 3 ("moderately active"), 912 in class 2 ("weakly active"), and 3470 in class 1 ("inactive"). For the binary analysis: 147 in class 4 ("active") and 4853 in class 1 ("inactive"). Since the data are experimental in nature, a certain number of false positives and false negatives are expected to occur. For the purposes of this study, no attempt was made to identify either, nor were corrections introduced to minimize the impact of these experimental errors. We intend to address this particular matter in a future paper.

We generated 1387 descriptors for each of the 20 986 members of the chemical library. First, 229 descriptors, distributed over the following categories, were calculated using the commercially available implementation of Cerius[2] (version 4.0; Molecular Simulations Inc., San Diego, CA):

fragment constants, conformational descriptors, electronic descriptors, graph-theoretic descriptors, topological descriptors, information-content descriptors, spatial descriptors, structural descriptors, and thermodynamic descriptors. Then 166 public ISIS MolsKeys were generated using ISIS/Host (version 3.0; MDL Information Systems Inc., San Leandro, CA), and 992 2D FingerPrints were generated using Unity (version 4.0; Tripos Inc., St. Louis, MO).

The DS training set was obtained using the diverse compound selection through a D-optimal design strategy (Euclidian distance metric, Tanimoto similarity coefficient, 10 000 Monte Carlo steps at 300 K, with a Monte Carlo seed of 11122, and termination after 1000 idle steps), as implemented in Cerius[2] (version 4.0; Molecular Simulations Inc., San Diego, CA).

The diverse selection of 5000 compounds (DS) was randomized with regard to the biological activity, yielding the diverse/randomized (DR) training set. To that purpose, 10 independent rounds of randomization were performed where compounds were randomly (using a random number generator) assigned to the activity bins proportionately to their initial distribution but without regard to their chemical structure and their measured biological activity.

RP is a method whereby a group of samples is recursively split at a branch point into two statistically distinct nodes. The statistical evaluation is performed using a Student's *t*-test. The data matrix consists of columns for each of the descriptors and rows for each of the samples in the training set. Each descriptor column is subjected to a process called splitting, in which the range (highest descriptor value − lowest descriptor value) is split into subranges. By systematically varying the splitting process, the statistical significance of each descriptor and its correlated range is determined. Branch points are identified by systematically evaluating the data matrix for the possibility to divide the matrix into statistically differentiated subsets based on their assigned category. The statistically most significant split then becomes a branch point in the RP tree. Each subset in the matrix is subsequently analyzed for further significant differentiation. The process ends when there are either no more significant splits to be obtained or when the minimum number of samples per node is reached. The program then proceeds to prune the tree to the appropriate tree depth as defined at the outset of the process. Sometimes, a molecule is included in a node because one of its descriptors increases the probability for it to be classified as "highly active". If this molecule, by virtue of its measured activity, belongs to a class other than the one to which it has been assigned, then that molecule is a "false positive" within that node. This at times occurs with a series of similar (congeneric) compounds. Conversely, molecules may have been eliminated from a node based on dissimilarity, but they should have been included. These molecules are "false negatives". Statistical models are generally geared to minimize both the number of false negatives and false positives. The default Gini splitting method is more susceptible to over training due to high node purity than parametric methods, but this has been at least partially addressed with the introduction of the Twoing splitting method that balances branches in RP trees.[29] The reader is kindly referred to the software documentation for a more detailed description of the RP process (Cerius[2], version 4.0; Molecular Simulations Inc., San Diego, CA). A method optimization/evaluation protocol was written that varied the RP conditions systematically. (The defaults as implemented in Cerius[2] are given in boldface.) The following conditions were considered: weighting by **Classes** (not varied), i.e., each class is considered of equal importance to the model rather than each compound; splitting method = Twoing/**Gini**/Greedy, i.e., the formalism that determines how groups are divided or partitioned into statistically distinct nodes or subgroups; maximum tree depth = 5/6/7/8/9/12/16/**20**, i.e., the maximum number of splits that may occur before the partitioning process terminates; pruning = **moderate** (not varied), i.e., the procedure that determines the appropriate statistically significant tree depth for each node; minimum number of samples per node = **1/100th** (not varied), i.e., a node or subgroup cannot contain fewer than 1% of all compounds in the training set; maximum number of knots per split = **20**−150 (in increments of 5), i.e., the maximum number of ways a descriptor range may be divided before statistical relevance is determined; and when applicable, number of cross-validation (XV) groups = 2, 3, 5, **10**, i.e., the number of groups used to test the statistical stability or significance of the model conditions. Therefore, any particular set of conditions can be characterized by "splitting method-maximum tree depth-maximum number of knots-number of XV groups (when applicable)".

Results and Discussion

General Definitions. There are two distinct measures for determining the success of an RP analysis: (1) "fold enrichment" and (2) "percent class correct" for the training set and the corresponding "percent hit recovery" for the test set.

(1) "Fold enrichment" represents the percentage of correctly predicted "hits" divided by the natural hit rate (the hit rate of the overall compound collection) expressed as a percentage, where the definition of "hit rate" is dependent on the class assignment. Data presented in this paper refer to class 4, i.e., "highly active" compounds, only. The optimization traces for the fold enrichment are presented in Figure 1.

(2) "Percent class correct" for the training set and the corresponding "percent hit recovery" for the test set are measures of the number of compounds correctly predicted to be "highly active" as a percentage of the total number of compounds known to be "highly active". The optimization traces for the percent hit recovery are represented in Figure 2.

Additionally, a "retrieval rate" can be defined. This is the number of compounds classified by the RP model as having an increased probability of being "highly active" expressed as a percentage of the total number of compounds under consideration in the test set.

Computational Methods Evaluation. A. Training Set Selection. The information content of the training set, whether a combinatorial library candidate for HTS or a

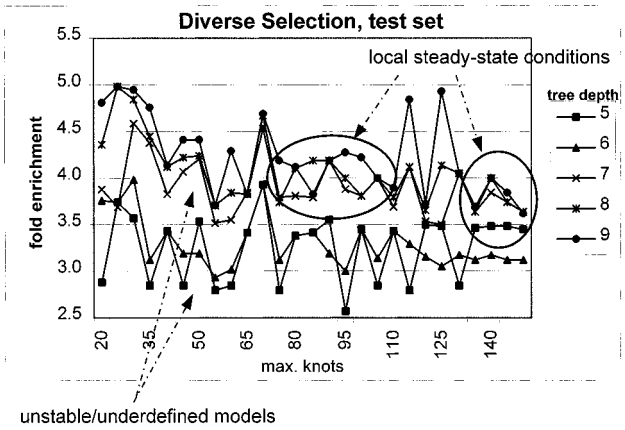


Figure 1. Optimization traces (fold enrichment). The fold enrichment obtained for the test set is plotted as a function of both the knot limit and the tree depth.

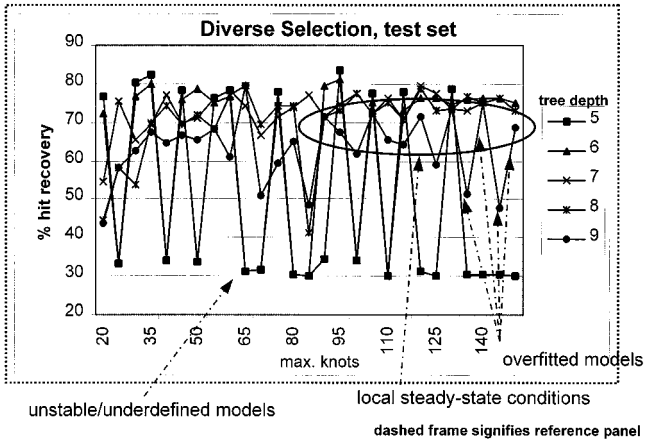


Figure 2. Optimization traces (percent hit recovery). The percent hit recovery obtained for the test set is plotted as a function of both the knot limit and the tree depth. The dotted frame signifies the reference panel for Figures 4 and 5.

statistical analysis data set, influences the efficiency and/or utility of the analysis methodology. For this reason different experimental design strategies have been developed for diverse compound selection from a larger chemical library or chemical diversity space.[33,34] The D-optimal design[33] strategy, as implemented in Cerius[2], was used to select our original 5000-member DS training set from the complete 20 986-member chemical library. In addition, we also generated a 5000-member random selection (RS) training set. Contrary to the DS training set, the RS training set is a stochastic sampling of the complete library and therefore represents the information content in proportion to its distribution in the complete library. In a sense, the information content is lower in the RS training set than in the DS training set because densely populated areas, with repetitive information, are sampled more frequently than sparsely populated areas, containing unique information, by random selection methods.

In one experiment, the RS training set (Twoing-8-90) predicted 5.7-fold enrichment, 60% class correct, and yielded 4.8-fold enrichment, 52% hit recovery (Table 1). The RS training set yielded even less predictive and unstable behavior at a tree depth of either 6 or 7 (Figure 5c). Although the fold enrichment, which reflects the density of the information matrix, compares favorably with the DS training set (4.2-fold when taken at Twoing-7-90, see Table 1), both the percent hit recovery and percent retrieval rate, which reflect the information content, are decreased. This probably is a reflection of the elimination of tentative false positives from the prioritization list. We intend to further investigate the impact of compound selection on the utility of RP, or other nonparametric methods, as applied to HTS data.

B. Method Optimization. The efficiency of RP can be expressed either as fold enrichment, or as percent class correct or percent hits retrieved for the training set and the test set, respectively. Ideally, the numbers for the training set and the test set match closely, i.e., the model shows good overall predictivity. But which value is the determinant factor for success?

One method, consensus scoring, emphasizes increases in hit rate by eliminating false positives from the prioritization list.[35] The aim of our analysis of HTS data is not simply enhanced hit rates, although it features prominently in the evaluation of the methodology. The aim, as we've chosen to define it, is (1) to increase the efficiency of our primary screens, i.e., increased hit rates; (2) to identify and pursue multiple chemotypes in order to develop compounds along parallel product lines, i.e., to achieve the highest percentage of chemotypes retrieved possible; and (3) the ability to explain nonlinear structure−activity relationships. Other factors such as the cost of a compound collection[36] may also contribute to the overall efficiency of the method, but they are not explicitly considered in this analysis.

Fold enrichment and percent hit recovery are not necessarily independent, rather they are interdependent. As the models become more sophisticated, e.g., increased tree depth, the activity is more narrowly defined, and as a result more false positives (compounds initially incorrectly included as active, but by a more refined model correctly identified as inactive) are eliminated from the model. However, concurrently, the method also eliminates more false negatives (compounds initially correctly identified as active, but subsequently incorrectly classified by the model as inactive), resulting in a better fold enrichment in the remaining models but a lower overall percent hit recovery (Figure 3).

Furthermore, we also considered the variability and reliability within the protocol: a low knot limit and small tree depth contribute to unstable behavior, whereas a high knot limit and large tree depth contribute to overfitting and add to computational expense (Figures 1 and 2). Therefore, we tried to identify the conditions that reliably and reproducibly yielded a model at an acceptable computational cost. The Twoing method (Figure 4a) balances the distribution of the branches of the tree, whereas the Gini method (Figure 4b) strives for the highest node purity.[29] When these two methods converge with regard to the tree depth, it can be argued that a suitable tree depth has been obtained. In the DS model, the maximal tree depth at which this occurs is 7 (Figure 2). Conversely, the Greedy method shows poor optimizability, and a low tree depth and knot limit will result in a rather poor predictive power of the model (Figure 4c).

To evaluate whether a method exhibits stable behavior or yields variable models due to perturbations we had to

Table 1. Selected Results from RP Models: Principal Output Measurements for Each of the Systematic Variations in the Training Set, and the Actualized Measurements for the Test Set

		training set		test set		
	protocol[a]	fold enrichment	% class correct	fold enrichment	% hit recovery	% retrieval rate
diverse selection	Gini-7-90	4.4	74	4.3	70	15
	Twoing-7-90	4.4	75	4.2	71	16
	Twoing-7-90-2	3.1	61	4.2	71	16
	Twoing-7-90-3	3.2	59	4.2	71	16
	Twoing-7-90-5	3.2	48	4.2	71	16
	Twoing-7-90-10	3.6	57	4.2	71	16
	Twoing-7-noknots	4.1	62	3.5	56	15
	Twoing-7-95	4.1	76	3.9	75	18
2500 diverse	Twoing-7-110	5.8	64	4.2	62	14
diverse/randomized	Twoing-7-90	1.8 ± 0.3	44 ± 16	0.9 ± 0.4	26 ± 14	27 ± 12
	Twoing-7-90-3	1.0 ± 0.1	28 ± 7	0.9 ± 0.4	26 ± 14	27 ± 12
random selection	Twoing-8-90	5.7	60	4.8	52	10
ISIS MolsKeys	Twoing-7-20	3.5	65	3.1	62	19
Unity FingerPrints	Twoing-7-20	4.1	73	3.5	67	18
binary analysis	Twoing-8-45	4.0	96	3.0	71	23
binary analysis + ISIS MolsKeys	Twoing-7-20	3.4	92	2.6	73	27

[a] The protocol is defined as "splitting method" − "tree depth" − "knot limit" − "number of XV groups".

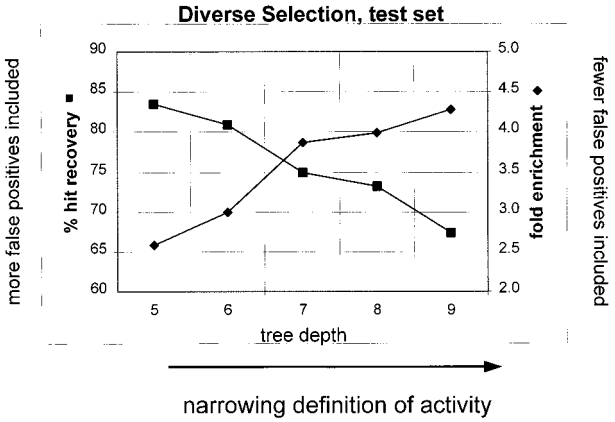


Figure 3. Interdependency of optimization values. The fold enrichment and percent hit recovery are (inversely) dependent on the tree depth.

establish a reference frame. One commonly employed reference frame is the running average of the centroid and its adjacent neighbors. The (first order) derived value "absolute value of the differential (running average − centroid)" then reflects the local variability of the function. If this value is close to zero, it indicates a "local steady-state". To differentiate between a "local steady-state" on an incline (all consecutive increment values are positive) or decline (all consecutive increment values are negative) of the optimization trace and one on a level region of the trajectory (consecutive increment values average out to zero), it becomes necessary to evaluate the "first derivative vs the knot limit" of the (second order) function. However, it is rather straightforward to inspect the curve visually and distinguish between the three possibilities. This then obviates the necessity to formalize this criterion.

If both the fold enrichment and the percent class correct converge on the "local steady-state" defined by a particular knot limit, then that knot limit is presumed to be the minimally acceptable knot limit. We have found it useful to evaluate a "local steady-state" on three consecutive differential values, i.e., a knots span of five consecutive steps. This is equivalent to three consecutive running averages and spans a total of 20 knots between the highest and the lowest conditions in the series. We found empirically that by defining a "local steady-state" (Figures 1 and 2) as variations of less than 0.1-fold enrichment and 2% class correct, we could eliminate many of the areas with irregularities. RP models selected with these criteria also tended to be more predictive for the test set, in both fold enrichment and percent hit recovery. These values are slightly more restrictive than, but in general agreement with, a standard deviation of 0.1-fold and 7% obtained during the randomization and cross-validation experiments (Table 1; Twoing-7-90-3).

The minimal knot limit in the DS optimization protocol at a maximal tree depth of 7 is therefore determined to be 90. The resulting values (Twoing-7-90) are 4.4-fold enrichment and 75% class correct for the training set, and 4.2-fold enrichment, a 71% hit recovery, and a 16% retrieval rate for the test set (Table 1). Since the data obtained from the test set are a relatively close reflection of the data from the training set, it is very likely that this approach is suited to select valid and predictive methods. Because the optimal conditions are inherently dependent on the training set, we expect that changes in the training set, such as changes in the chemical composition or in the biological data, such as a different target selection, will require reoptimization of the RP conditions.

In addition, we also evaluated the built-in autoselection protocol, i.e., the "no knot limit" setting, which yielded the following data: Twoing-7-noknots predicted 4.1-fold enrichment and 62% class correct, and yielded 3.5-fold enrichment, 56% hit recovery, and a 15% retrieval rate (Table 1). The discrepancy between our optimal conditions and those selected by the program probably find its roots in the undisclosed optimization criteria of this particular implementation. As a result, the Twoing-7-noknots protocol has a lower predictive capability for this data set than the models with manually and empirically determined optimal conditions.

C. Randomization and Cross-Validation Experiments. When evaluating the efficiency of any methodology, one has to take into account how predictive the method developed

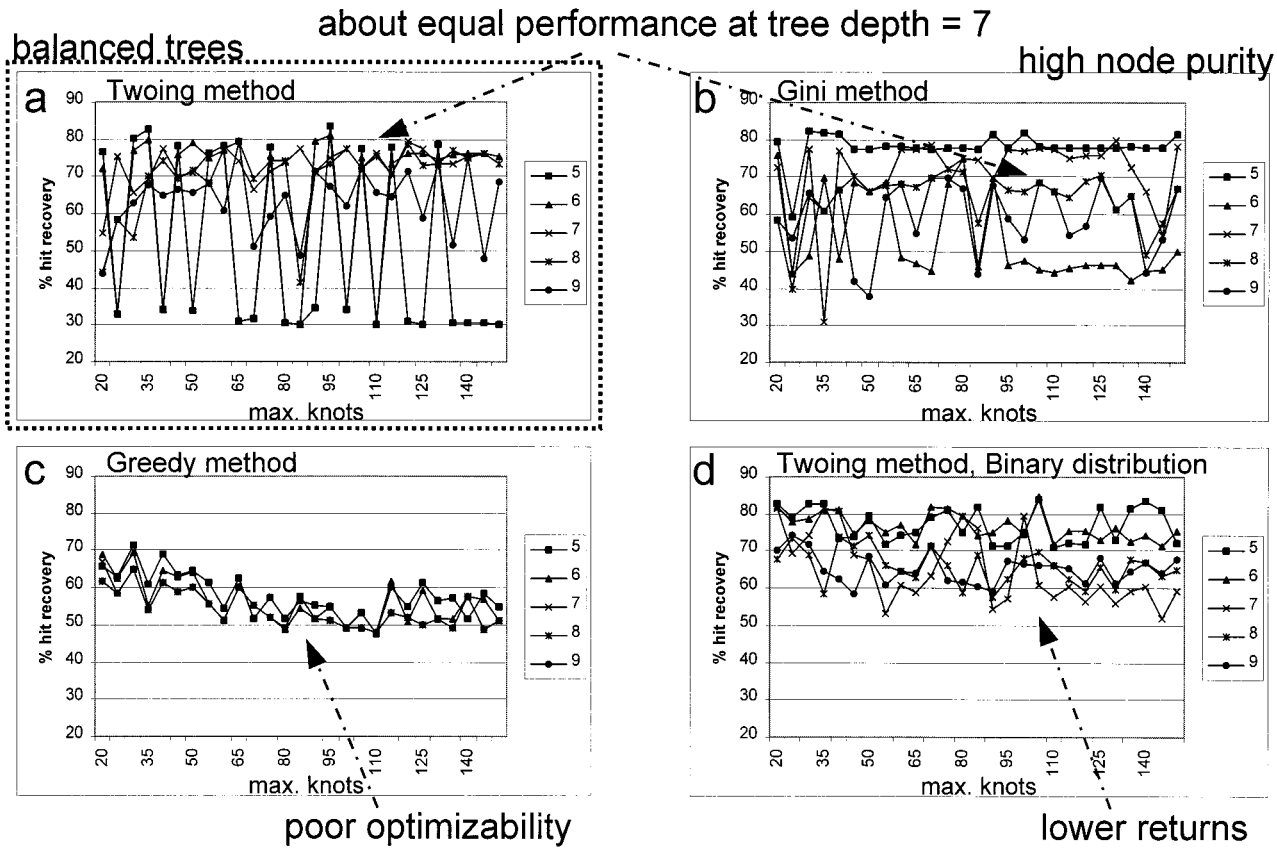


Figure 4. Comparison between protocols. The percent hit recovery obtained for the test set is plotted as a function of both the knot limit and the tree depth. (a) Application of the Twoing splitting method to the DS data set. (b) Application of the Gini splitting method to the DS data set. (c) Application of the Greedy splitting method to the DS data set. (d) Application of the Twoing splitting method to a binary distribution of the DS data set.

on the training set is when applied to the test set. In parametric methods this usually is quantitated in the form of correlation coefficients and cross-validation values. Due to the noncorrelative nature of nonparametric methods in general, regression has so far been inpracticable. Recently, a regression method for nonparametric evaluation of small data sets has become available.[37] The validation method implemented in Cerius[2] is cross-validation. In addition to the cross-validation experiments described below, we performed 10 independent randomization trials to remove selection bias.

The results of the randomization experiment ($n = 10$; Twoing-7-90) are presented in Table 1. Under these conditions, RP apparently overstates its efficiency with regard to a fully randomized training set. This may be a result of the distribution of chemotypes in the training and test sets, but it cannot be unambiguously proven. More importantly, cross-validation of the DR training set (Twoing-7-90-3) yielded results that are in good agreement with results obtained with the test set. This further supports the notion that there is a bias present in the training set that is not present in the test set. The only difference in composition between the test set and the training set is a mathematically introduced one. The mathematical process that introduced this bias must have been the diverse selection process that separated the training set from the test set. It is therefore intuitive, but not factually proven, that the measured bias is a result of the distribution of chemotypes between training and test sets.

The cross-validation (XV) experiment led us to investigate how the "information content" of the training set influences the outcome of the analysis. We found that at a low number of XV groups (2 or 3), i.e., high information dilution, the predictivity of the models fell short of the expectations based on a larger number of XV groups (5 or 10). When the XV experiment was run with five XV groups, i.e., 80% of the training set, the model values of the training set and the test set were in good agreement (Table 1). Alternatively, when two XV groups were used, i.e., 50% of the training set, the XV model was less predictive of the full model. A similar effect is seen when the full training set is reduced to 2500 diversely selected compounds from the original 5000-member DS training set. Twoing-7-110 predicted 5.8-fold enrichment, and yielded 4.2-fold enrichment (Table 1), but with rather unstable optimization traces (not shown) and a significant discrepancy between predicted and realized yields. This less predictive model reflects the loss of information content in the training set selection, and deserves a closer examination.

This last point is also illustrated with the results presented in Figure 5. Here, we changed the binning scheme to be more restrictive when assigning compounds to the more active classes (Table 4). This had two intended effects: (1) it decreased the initial hit rate, thereby allowing us to focus on the more potent hits only; and (2) it diluted the information content available to the RP algorithm. Rather surprisingly, we found that models based on these more

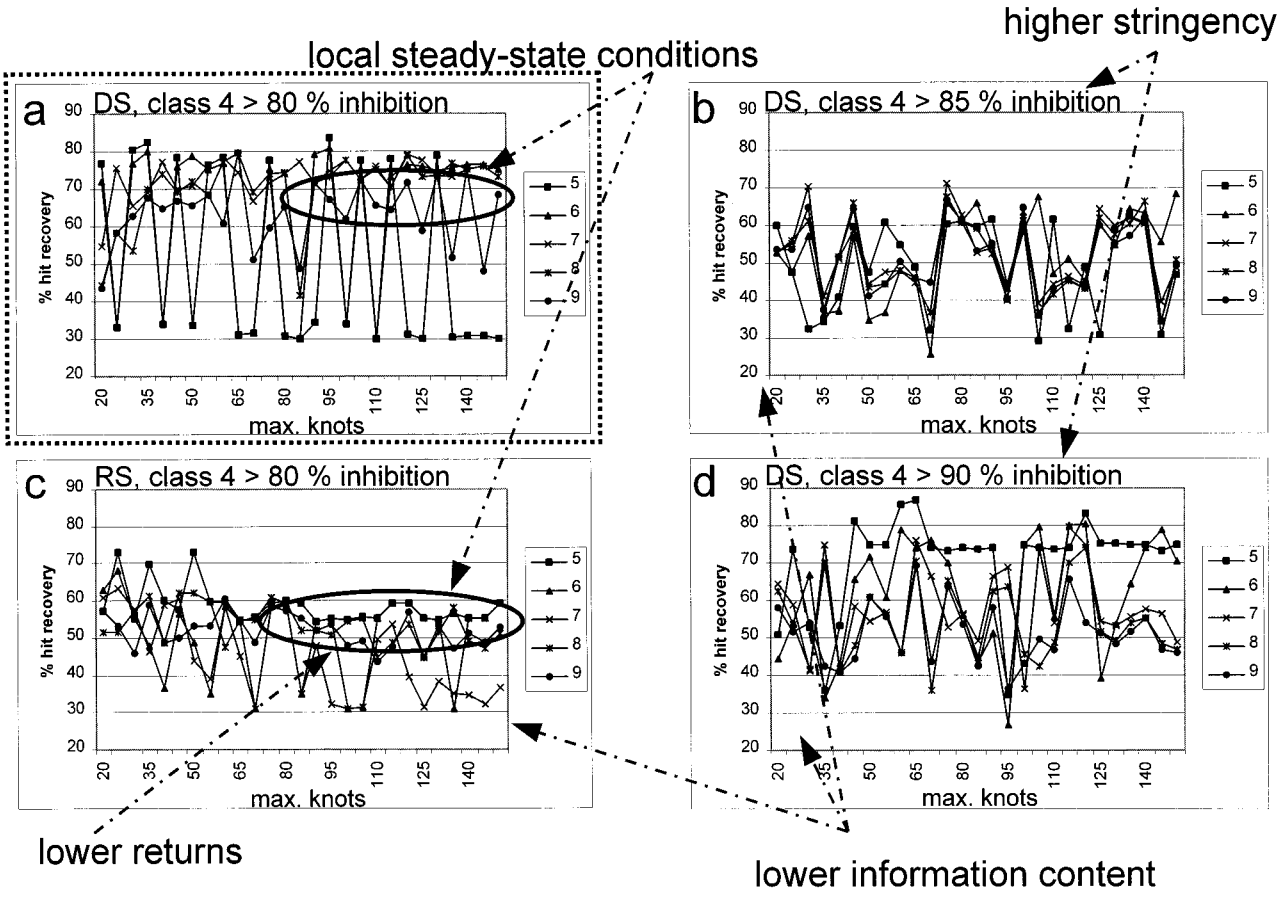


Figure 5. Comparison between training sets. The percent hit recovery obtained for the test set is plotted as a function of both the knot limit and the tree depth. (a) Use of an 80% threshold value applied to the DS data set. (b) Use of an 85% threshold value applied to the DS data set. (c) Use of an 80% threshold value applied to the RS data set. (d) Use of a 90% threshold value applied to the DS data set.

restrictive binning schemes were less stable, less predictive, and had an overall lower yield (Figure 5b,d). However, these results are entirely consistent with the findings from our binary analysis, as discussed below. It probably finds its origin in the fact that prediction accuracy is significantly compromised near any artificial threshold, such as represented by the binning schemes in Table 4.

Experimental Results. A. Overall Results. When the criteria and considerations detailed above (Computational Methods Evaluation section) were applied to the 5000-member DS training set, we found that the models converged on a tree depth of 7 (Figures 1, 2, and 4). The resulting preferred RP model (Twoing-7-95) yielded a 3.9-fold enrichment and a 75% class correct for the test set (Table 1). The model predicted that 18% of the compounds in the test set belong to the "highly active" class, i.e., a reduction in the number of compounds to be screened by 82%. This 18% retrieval rate should recover more than 18% of the "highly actives" present in the test set (if hits were proportionally distributed between the 18% selected and the 82% remaining compounds) in order to be deemed successful. Indeed, in using this model we retrieved 75% of all "highly active" compounds present, thereby enhancing our hit rate some 4-fold. This result satisfies one of the criteria laid out in the Introduction: the ability to increase the efficiency of our primary screens. There currently are two limitations to this method that require further attention. The first limitation is the percent hit recovery. This directly impacts our ability to satisfy our second condition: the option to pursue multiple chemotypes in order to develop compounds along parallel product lines. This aspect of the analysis is addressed immediately below (Chemotypes and Nodes section). The second limitation is the percent retrieval rate. If we consistently retrieve about 20% of the total test set, than we cannot attain better than 5-fold improvement using this method. We know, since this is a retrospective analysis, that the hit rate in both the training and the test set is well below 20%. In fact, the hit rate is on the order of 3%, which means that the maximal theoretical improvement is about 30-fold. We are currently investigating how the false positives in an HTS data set can be identified and subsequently eliminated from the "highly active" category. Presumably, this should reduce the interference of false positives with the classification and thereby reduce the number of compounds erroneously predicted as "highly active". This would in turn reduce the number of compounds retrieved, and concurrently decrease the percent retrieval rate and increase the efficiency of the methodology.

B. Chemotypes and Nodes. As stipulated earlier in this paper, the distribution of chemotypes within the compound collection may play a role in the performance of the RP models. This directly impacts our wish to pursue multiple chemotypes at the same time in order to develop compounds along parallel product lines. Sometimes compounds can be used for different indications, such as gastrointestinal versus central nervous system diseases. At other times, it can be

Table 2. Results per Terminal Node (DS80: Twoing-7-90): Distribution of Each of the Compounds Assigned to Class 4 with Respect to Their Placement in the Terminal Nodes I−VIII

	hits		class 4		hit rate (%)		fold	
node	train[a]	test	train	test	train	test	train	test
I	12	38	107	367	11.2	10.4	3.8	3.5
II	4	9	84	242	4.8	3.7	1.6	1.3
III	4	8	52	147	7.7	5.4	2.6	1.9
IV	4	8	51	207	7.8	3.9	2.7	1.3
V	30	90	103	304	29.1	29.6	9.9	10.1
VI	46	136	379	1186	12.1	11.5	4.1	3.9
VII	5	2	50	153	10.0	1.3	3.4	0.4
VIII	14	25	147	487	9.5	5.1	3.2	1.8

[a] train = training.

quite useful to have one lead compound progressing toward the clinic while another one serves as a so-called "back-up" compound. After all, xenobiotics are frequently not readily absorbed, and can be extensively metabolized and excreted.

In an RP model each terminal node represents a different stratification of the data that is not necessarily analogous to, or even consistent with, another node. This opens up the possibility that different nodes may represent differences either in chemical or in biological stratification. We therefore investigated the results for each of the terminal nodes individually. On the basis of a general definition of chemical core structures derived from the combinatorial synthetic process, eight distinct chemotypes could be identified within the training and test sets (CT1 through CT8).

In Table 2, we have collected data for the terminal nodes in the DS/Twoing-7-90 RP model. It is apparent that there is "significant" variability between the nodes. This may indicate the presence of distinct "binding modes" or allosterism in the data set. Whereas some nodes, e.g., node V, show robust (about 10-fold) increases in fold enrichment in both the training and the test set, other nodes, e.g., node II and III, do not perform as well. Moreover, the results for node VII completely miss the mark, which may merely be a reflection of the small number of hits in the training set (five) and the test set (two). In contrast, the results obtained for node VI reflect the overall results, because of the large number of compounds (1186) assigned to that node.

Differentiation by chemotype of the terminal nodes (Table 3) indicates that all chemotypes representing "hits" are correctly identified by this method. Although only about 70% of all "hits" are retrieved using this RP model, it is gratifying to see that a full complement of chemotypes has been identified. This then leads us to believe that we have successfully established our second goal, which is the option to pursue multiple chemotypes with demonstrated activity against a single biological target. Again, we must keep in mind that the set of compounds identified by the HTS assay as "highly active" may contain some false positives and equally well may have failed to identify false negatives. The statistical methods could have left out false positives (by chance or by failure to conform to the model) and may have included some of the false negatives. This would be reflected in an overall lower percent hit recovery and a higher percent retrieval rate. Until we design an experiment that will unambiguously identify false positives and false negatives, we can only surmise that the overall impact of falsely identified compounds is negligible. Whereas it may be possible to identify false positives from an HTS assay by performing some sort of secondary assay, due to the sheer numbers involved it will be virtually impossible to address the matter of false negatives in a systematic fashion. We are planning to evaluate the impact of identification of false positives on RP models soon.

There seems to be a preponderance of a particular chemotype (CT7) in nodes II, III, IV, VI, and VIII, well above the prevalence in the overall distribution. Moreover, this chemotype is lacking in nodes V (which comprises mainly CT1 and CT3) and VII (in majority CT6), whereas CT4 is only present in node VIII. These results support the notion that at least three, if not more, different "binding modes" may be represented and identified by RP analysis of this data set.

Cross-correlation of terminal node and chemotype contributions demonstrates that node V, which lacks CT7 but contains a majority of CT3, yields the highest fold enrichment: 10-fold. Conversely, node III consisting of 100% CT7 yields a below average result: 1.9-fold enrichment. These results are indicative of a nonlinear structure−activity relationship within this data set.

This, then, brings us back to the earlier supposition that the HTS data may not be normally distributed. As can be seen in Figure 6, the HTS data (plotted points) do not follow a strictly Gaussian behavior (fitted line). Rather, the HTS data have a higher than normal incidence in the 30th−50th percentile range and a lower than normal incidence at the higher than 70th percentile range. Nevertheless, the central tenet of the central limit theorem is that data sets will appear to be normally distributed as long as the sample size is large enough. At a sample size of over 20 000 data points, our data set certainly has the resemblance of being normally distributed. It does, however, raise the question of whether a collection of multimodal or multiple binding site models could be hidden within this distribution. The results from the RP analyses suggest that the latter is indeed the case, but it cannot be unambiguously proven or disproven based on the distribution data alone.

C. Alternate Descriptor Sets. Whereas to the medicinal chemist it may be obvious that 3D physicochemical or stereochemical information is as important a determinant of biological activity as the chemical composition of a compound, computational chemical methods have focused mainly on describing chemical (diversity) space in two dimensions, e.g., MDL MolsKeys[38] and 2D FingerPrints,[39] to facilitate throughput and ease of calculation (no geometry optimization, and conformer analysis are required for 2D descriptors). Recently, progress has been made to describe compounds in terms of their 3D information content, such as pharmacophore definition triplets,[39] or a combination of 2D and 3D descriptors such as those implemented in CODESSA[40] (comprehensive descriptors for structural and statistical analysis). Electrotopological descriptors, as represented by the E-state keys of Kier and Hall[41] and implemented in Cerius[2], try to incorporate 2D as well as 3D information by describing the chemical connectivity (topology) of a molecule.

Table 3. Distribution of Chemotypes per Terminal Node (DS80: Twoing-7-90): Relative Distribution of Each of the Chemotypes (CT1−CT8) with Respect to Their Occurrence in the Terminal Nodes I−VIII

	% per node								% overall		
	predicted									actual	
	I	II	III	IV	V	VI	VII	VIII	class 4	hits	library
CT1	5.3	23.0	0.0	12.8	22.2	15.5	14.8	0.3	12.4	23.9	10.5
CT2	13.9	4.9	0.0	0.0	0.2	0.9	0.0	15.5	4.7	8.9	6.9
CT3	32.7	5.8	0.0	0.0	58.5	5.7	5.4	0.3	12.7	26.1	7.5
CT4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	20.0	3.0	1.5	9.1
CT5	9.1	15.6	0.0	5.0	13.3	0.9	14.8	0.8	5.2	3.3	6.6
CT6	2.1	0.6	0.0	1.6	5.7	0.1	65.0	0.2	4.3	2.2	7.4
CT7	36.9	50.0	100.0	80.6	0.0	76.9	0.0	56.2	56.7	31.8	33.4
CT8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	6.7	1.0	2.2	18.7

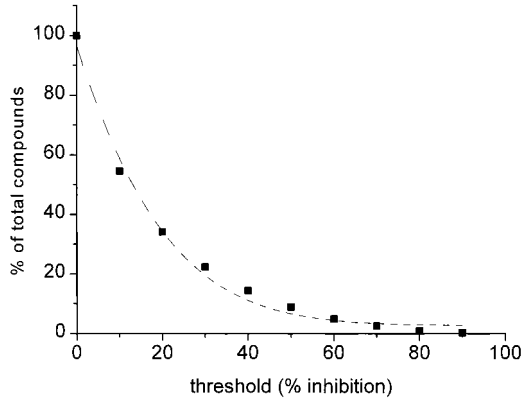


Figure 6. Distribution of HTS data. The relative distribution of biological data as recorded by the HTS assay expressed by decile (squares), and a fitted Gaussian distribution function (dotted line).

Recently, Dixon and Villar[42] reported on their efforts to distinguish three different pharmacological classes from those present in the CMC,[10] primarily based on similarity measures. Since they demonstrated superiority of the ISIS MolsKeys over Molconn-X descriptors, we investigated the impact of descriptor set selection on the outcome of our RP analysis.

We designed an experiment where the 166 public ISIS MolsKeys were represented in binary form (0 defines the absence, and 1 the presence of a particular feature), and found that RP (Figure 5a,b; Twoing-7-20) predicted a 3.5-fold enrichment and 65% class correct and yielded a 3.1-fold enrichment, with a 62% hit recovery and a 19% retrieval rate (Table 1). Whereas Dixon and Villar[42] were able to correctly identify close to 90% of the actives in a H_2 receptor antagonist data set by examining 20% of the compounds and demonstrated a 4-fold enhancement over the random sampling method, we were unable to achieve better results with the ISIS MolsKeys than with our original descriptor set. This probably reflects the presence of descriptors in our descriptor set other than the substructurally defined ISIS MolsKeys.

In addition, we designed an experiment based on 992 bitkeys derived from the Unity 2D FingerPrints (2DFP). Under these conditions (Figure 5a,b; Twoing-7-20), the algorithm predicted a 4.1-fold enrichment and 73% class correct and yielded a 3.5-fold enrichment, with a 67% hit recovery and a 18% retrieval rate (Table 1). Likewise, this probably reflects the absence of physicochemical (whole molecule) and 3D descriptors required for optimal performance by this data set.

Table 4. HTS Data Binning Schemes. Compounds in the Training Sets Were Assigned to One of Four Activity Bins, Depending on Their Biological Activity as Recorded by the HTS Assay. Class 4 Is Considered "Highly Active", Class 3 "Moderately Active", Class 2 "Weakly Active", and Class 1 "Inactive". The Distributions for Three Different Thresholds Are Presented Here

class	"80"[a]	"85"[a]	"90"[a]
4	≥80 (2.9%)	≥85 (2.1%)	≥90 (1.2%)
3	<80, ≥50 (9.4%)	<85, ≥60 (6.0%)	<90, ≥70 (3.5%)
2	<50, ≥25 (18.2%)	<60, ≥30 (17.7%)	<70, ≥30 (21.1%)
1	<25 (69.4%)	<30 (74.2%)	<30 (74.2%)

[a] Numbers in parentheses represent the classification rates for the DS training set.

D. Binary Analysis. Gao and Bajorath[43] reported that an increase in accuracy from 84% for 2D QSAR to 94% could be obtained using binary QSAR. We found that RP (Twoing-8-45; Figure 4d) based on a binary distribution decreased both the accuracy (from 75 to 71% hit recovery) and the efficiency (from 3.9- to 3.0-fold) of the models. This reflects a decrease in predictivity of the model rather than an improvement of the training set model and also results in unstable optimization traces. We therefore speculate that the "fuzzy assignment" approach that we have employed, i.e., four activity classes rather than just two, allows the algorithm to compensate for false positive and false negative assignments, without compromising the node purity. A strictly binary classification forces the algorithm to apply penalties to, e.g., compounds having data that fall within a class 3 classification, but which the model assigned to class 4, whereas the distinction in HTS data between "highly active" and "moderately active" is not necessarily that clear (Table 4) and possibly within the statistical confidence interval. This hypothesis is further supported by the finding of Gao and Bajorath that the prediction accuracy was significantly compromised (about 60% accuracy) near the binary threshold.[43] This problem is exacerbated in the case of percent inhibition data, such as associated with our HTS data set, where the threshold is usually set at the edge of the upper confidence interval, resulting in overlap of the "active" and "inactive" categories whereby most "active" compounds (within the uncertainty) could equally well be placed in the "inactive" category, but only a limited number of "inactive" compounds qualify to be placed in the "active" category. The problem is also compounded by the ceiling effect imposed on the data set by setting a 100% limit as the maximal response, which restricts all compounds passing the

threshold to a class 4 assignment, thereby potentially bringing them within the confidence interval of class 3 in a quaternary classification or the "inactive" class of a binary classification.

We also investigated whether a binary descriptor set would be more appropriate for a binary stratification of biological data by applying the binary classification to the 166 public ISIS MolsKeys, and we found this to not appreciably improve the predictivity of the models (Table 1; Twoing-7-20), nor affect the quality of the models (Figure 5a,b). We therefore based all work on a quaternary classification of the biological data.

Concluding Remarks

In this paper we have demonstrated that nonparametric methods with nonbinary (continuous range) descriptors derived from a 20 986-member combinatorial library can be effectively employed to differentiate between active and inactive compounds, based on data from an experimental HTS assay. Moreover, in a truly experimental fashion we have demonstrated that we can improve the hit rate of our primary screens by about 4-fold, and in doing so correctly identify 75% of all hits while reducing the size of the chemical library to be screened by over 80%. Furthermore, even though up to 25% of all individual hits go undetected when this particular analysis is employed as a prescreening method, all chemotypes with known activity were correctly identified. This then opens up the possibility to pursue missed hits and potentially identify false negatives during subsequent screening or SAR development.

Thus, we believe that we have satisfied the goals set at the outset of the analysis:

(1) The ability to increase the efficiency of our primary screens: we have demonstrated that this approach allows us to increase the efficiency of this particular HTS assay by a moderate 4-fold and speculate that we can achieve higher efficiency still by fine-tuning the process and HTS parameters.

(2) The option to pursue multiple chemotypes in order to develop compounds along parallel product lines: this approach has allowed us to correctly identify almost 75% of all hits, and even more importantly, all chemotypes known to be active.

(3) The ability to explain nonlinear structure−activity relationships: the analysis indicates that at least two and possibly more binding modes are represented by the combinatorial library and HTS data set, thus allowing us to pursue different pathways to treat ion channel related diseases.

While the role of molecular diversity and the influence of false positive data on interpretation of HTS screening results has been the subject of much speculation, most computational methods described to date utilize confirmed data from compound collections that tend to be poorly diverse. In reality, the level of diversity in a screening set can be highly controlled. On the other hand, HTS data by its nature is unconfirmed and will contain some level of false positive and false negative data. One of the goals of our work is to develop a method that is sufficiently robust to accommodate false positives and false negatives without compromising the utility of the results. Our current studies using "real world" HTS screening data show that RP will accommodate this. We are now investigating the role of library diversity on this process. The results of this study will be the subject of a future report.

On the basis of the tentative evidence that RP can differentiate between multiple binding modes, we speculate that it will be possible to include multiple targets in a single HTS data set to deduce SARs for the individual targets. Additionally, we surmise that if the biological targets are related by their gene family, we may be able to establish a generalized SAR for the entire gene family. We have dubbed this approach "gene family libraries". Chemical libraries thus developed should have a higher propensity to identify hits from gene family related HTS data, even if the individual target has not been screened before. Furthermore, this approach could allow us to "stake out" the territory in diversity space where chemical and biological diversity spaces intersect. This work is ongoing and will be reported in the future.

References and Notes

(1) Delaney, J. S. Assessing The Ability of Chemical Similarity Measures to Discriminate Between Active and Inactive Compounds. *Mol. Diversity* **1995**, *1*, 217−222.
(2) Bayley, M. J.; Willett, P. Binning Schemes for Partition-Based Compound Selection. *J. Mol. Graphics Modell.* **1999**, *17*, 10−18.
(3) Wang, J.; Ramnarayan, K. Toward Designing Drug-Like Libraries: A Novel Computational Approach for Prediction of Drug Feasibility of Compounds. *J. Comb. Chem.* **1999**, *1*, 524−533.
(4) Klopman, G. The MultiCASE Program II. Baseline Activity Identification Algorithm (BAIA). *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 78−81.
(5) Burden, F. R. Holographic Neural Networks as Nonlinear Discriminants for Chemical Applications. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 47−53.
(6) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.
(7) Magalhães, N. S. S.; Cabral De Holanda Cavalcanti, S.; Alencar De Menezes, I. R.; Antunes De Sousa Araújo, A.; Magalhães De Oliveira, H.; Alves, A. J. Automated Search for Potentially Active Compounds by Using Cluster Trees. *Eur. J. Med. Chem.* **1999**, *34*, 83−92.
(8) Chen, X.; Rusinko, A.; Tropsha, A.; Young, S. S. Automated Pharmacophore Identification for Large Chemical Data Sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 887−896.
(9) Available Chemical Directory. MDL Information Systems Inc.: San Leandro, CA.
(10) Current Medicinal Chemistry. MDL Information Systems Inc.: San Leandro, CA.
(11) World Drug Index. Derwent Inc.: Vienna, VA.
(12) Coffen, D. L.; Baldino, C. M.; Lange, M.; Tilton, R. F.; Tu, C. Molecular Diversity, Biological Activity and Common Ground Shared by Both. *Med. Chem. Res.* **1998**, *8*, 206−218.
(13) Lucic, B.; Trinajstic, N. Multivariate Regression Outperforms Several Robust Architectures of Neural Networks in QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 121−132.
(14) Fujita, T.; Hansch, C. Analysis of the Structure−Activity Relationship of the Sulfonamide Drugs Using Substituent Constants. *J. Med. Chem.* **1967**, *10*, 991−1000.
(15) Hunt, P. A. QSAR Using 2D descriptors and TRIPOS' SIMCA. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 453−467.

(16) Rogers, D.; Hopfinger, A. J. Application of Genetic Function Approximation to Quantitative Structure−Activity Relationships and Quantitative Structure−Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854−866.
(17) Dohoo, I. R.; Ducrot, C.; Fourichon, C.; Donald, A.; Hurnik, D. An Overview of Techniques for Dealing with Large Numbers of Independent Variables in Epidemiologic Studies. *Prev. Vet. Med.* **1997**, *29*, 221−239.
(18) Holzgrabe, U.; Mohr, K. Allosteric Modulators of Ligand Binding to Muscarinic Acetylcholine Receptors. *Drug Discovery Today* **1998**, *5*, 214−222.
(19) Zwart, R.; Vijverberg, H. P. Potentiation and Inhibition of Neuronal Nicotinic Receptors by Atropine: Competitive and Noncompetitive Effects. *Mol. Pharmacol.* **1997**, *52*, 886−895.
(20) Chen, H. S.; Liptin, S. A. Mechanism of Memantine Block of NMDA-activated Channels in Rat Retinal Ganglion Cells: Uncompetitive Antagonism. *J. Physiol.* **1997**, *499* (Pt 1), 27−46.
(21) Nelson, L. M.; Bloch, D. A.; Longstreth, W. T.; Shi, H. Recursive Partitioning for the Identification of Disease Risk Subgroups: A Case-Control Study of Subarachnoid Hemorrhage. *J. Clin. Epidemiol.* **1998**, *51*, 199−209.
(22) Cook, E. F.; Goldman, L. Empiric Comparison of Multivariate Analytic Techniques: Advantages and Disadvantages of Recursive Partitioning Analysis. *J. Chron. Dis.* **1984**, *37*, 721−731.
(23) Hand, D. J. Statistical Methods in Diagnosis. *Stat. Methods Med. Res.* **1992**, *1*, 49−67.
(24) Scott, C. B.; Scarantino, C.; Urtasun, R.; Movsas, B.; Jones, C. U.; Simpson, J. R.; Fischbach, A. J.; Curran, W. J. Validation and Predictive Power of Radiation Therapy Oncology Group (RTOG) Recursive Partitioning Analysis Classes for Malignant Glioma Patients: A Report Using RTOG 90-06. *Int. J. Radiat. Oncol., Biol., Phys.* **1998**, *40*, 51−55.
(25) Lacher, D. A. Comparison of Non-parametric Recursive Partitioning to Parametric Discriminant Analyses in Laboratory Differentiation of Hypercalcemia. *Clin. Chim. Acta* **1991**, *204*, 199−208.
(26) Dobbertin, M.; Biging, G. S. Using the nonparametric classifier CART to model forest tree mortality. *Forest Sci.* **1998**, *44*, 507−516.
(27) Topliss, J. G. Utilization of Operational Schemes for Analog Synthesis in Drug Design. *J. Med. Chem.* **1972**, *15*, 1006−1011.
(28) Topliss, J. G. A Manual Method for Applying the Hansch Approach to Drug Design. *J. Med. Chem.* **1977**, *20*, 463−469.
(29) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. Classification and Regression Trees. Wadsworth, 1984.
(30) Young, S. S.; Hawkins, D. M. Analysis of a 2^9 Full Factorial Chemical Library. *J. Med. Chem.* **1995**, *38*, 2784−2788.
(31) Rusinko, A. R.; 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.
(32) MDL Drug Data Report. MDL Information Systems Inc.: San Leandro, CA.
(33) Hassan, M.; Bielawski, J. P.; Hempel, J. C.; Waldman, M. Optimization and Visualization of Molecular Diversity of Combinatorial Libraries. *Mol. Diversity* **1996**, *2*, 64−74.
(34) Higgs, R. E.; Bemis, K. G.; Watson, I. A.; Wikel, J. H. Experimental Designs for Selecting Molecules from Large Chemical Databases. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 861−870.
(35) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus Scoring: A Method for Obtaining Improved Hit Rates from Docking Databases of Three-Dimensional Structures into Proteins. *J. Med. Chem.* **1999**, *42*, 5100−5109.
(36) Young, S. S.; Sheffield, C. F.; Farmen, M. Optimum Utilization of a Compound Collection or Chemical Library for Drug Discovery. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 892−899.
(37) Constans, P.; Hirst, J. D. Nonparametric Regression Applied to Quantitative Structure−Activity Relationships. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 452−459.
(38) McGregor, M. J.; Pallai, P. V. Clustering Large Databases of Compounds: Using the MDL "Keys" as Structural Descriptors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 443−448.
(39) Matter, H.; Pötter, T. Comparing 3D Pharmacophore Triplets and 2D Fingerprints for Selecting Diverse Compound Subsets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1211−1225.
(40) Menziani, M. C.; Montorsi, M.; De Benedetti, P. G.; Karelson, M. Relevance of Theoretical Molecular Descriptors in Quantitative Structure−Activity Relationship Analysis of α_1-adrenergic receptor antagonists. *Bioorg. Med. Chem.* **1999**, *7*, 2437−2451.
(41) Kier, L. B.; Hall, L. H. *Molecular Structure Description: The Electrotopological State*; Academic Press: New York, 1999.
(42) Dixon, S. L.; Villar, H. O. Investigation of Classification Methods for the Prediction of Activity in Diverse Chemical Libraries. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 533−545.
(43) Gao, H.; Bajorath, J. Comparison of Binary and 2D QSAR Analyses Using Inhibitors of Human Carbonic Anhydrase II as a Test Case. *Mol. Diversity* **1999**, *4*, 115−130.

CC0000747