Use of Recursion Forests in the Sequential Screening Process: Consensus Selection by Multiple Recursion Trees

A. Michiel van Rhee*

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

Received February 4, 2003

The application of Cheminformatics to High-Throughput Screening (HTS) data requires the use of robust modeling methods. Robust models must be able to accommodate false positive and false negative data yet retain good explanatory and predictive power. Recursive Partitioning has been shown to accommodate false positive and false negative data in the model building phase but suffers from a high false positive rate in the prediction phase, especially with sparse data sets such as HTS data. Here, we introduce Consensus Selection as a procedure to decrease the false positive rate of Recursive Partitioning-based models. Consensus Selection by Multiple Recursion Trees can increase the hit rate of a High-Throughput Screen in excess of 30-fold while significantly reducing the false positive rate relative to single Recursion Tree models.

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 (typically much less than 1%), resulting in very sparse data sets. Our interest in discovering novel, small molecule modulators (inhibitors, activators, 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. Moreover, it has led us to explore and implement a sequential screening process (see also ref 1).

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. The limitation of most Quantitative Structure−Activity Relationship (QSAR) methods is that a single (quasi-) linear equation is presumed to account for all biological activity, which is presumed to reside in a single binding site. 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[2−4] 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+}-sensitive channels), are noncompetitive, or uncompetitive, allosteric modulators. It has been shown that this problem can be addressed using Probabilistic Structure−Activity Relationship (PSAR) models based on Recursive Partitioning (RP).[5,6] Whereas the selection of variables in parametric methods is determined by their impact on correlation, RP focuses on classification. RP is a nonparametric method, capable of using factorial (categorical), integer (discrete), and decimal (continuous) descriptors. All variables are considered simultaneously during every step of the model-building phase. Moreover, no particular statistical distribution, dependency, or normality is assumed for any of the variables.[7] 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 of observations 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.[8]

RP is a method whereby a group of samples is recursively split at a branch point into two statistically distinct nodes. 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 is split into subranges. By systematically varying the splitting process, the most effective splitting point for each descriptor 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 (activity) category. The 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 either when there are 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

* Corresponding author phone: (919)941-5206; fax: (919)941-0813; e-mail: mvanrhee@icagen.com.

10.1021/ci034023j CCC: $25.00
Published on Web 05/02/2003

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 should have been included. These molecules are "false negatives". Statistical models are generally geared to minimize only one type of error. Selection of the splitting rule has considerable consequences for the outcome. The Gini splitting method, which is the default in the commercial release of RP by Accelrys, was designed to isolate the largest segment of the data set in the least number of steps, i.e. it favors high node purity. In highly unbalanced data sets such as HTS data, this could easily lead to all compounds being classified as "inactive" (see Results and Discussion section). The Twoing splitting method, used in this study, was designed to split the data set into more or less equal parts, i.e. maintain parity, thus trying to balance the impact of both true positives and true negatives.[7]

RP was demonstrated by Young and Hawkins,[9,10] as early as 1995, to be a powerful method of harnessing the information content of a 512-member combinatorial chemical library. Whereas a typical (Q)SAR series usually comprises as few as 20 or as many as 50 compounds, their approach increased the dimension of the problem at least by an order of magnitude. We have previously demonstrated that RP can also be effectively employed when assessing HTS data generated against a 20 000-member noncombinatorially derived diversified chemical library.[5]

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.[11] On the one hand, the level of diversity in a screening set can be highly controlled. On the other hand, HTS data by their nature are 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 models. Here we aim to address this issue with the introduction of Consensus Selection by Multiple RP Trees. In this study, we contrast this method with tuning of the minimum node size as an alternative way to decrease the false positive rate of RP models.

METHODS

Standardized 3D starting geometries were obtained for all compounds using the UNITY (version 4.2; Tripos Inc., St. Louis, MO) dbtranslate utility in conjunction with the CONCORD (version 4.04; developed by R. S. Pearlman, A. Rusinko, J. M. Skell, and R. Baducci at the University of Texas, Austin, and distributed by Tripos Inc., St. Louis, MO) program. Three hundred eighty-three descriptors were calculated using Cerius^2 (version 4.5; Accelrys Inc., San Diego, CA; selected from the following categories: E-state keys, Electronic, Information Content, Molecular Shape Analysis, Spatial, Structural, Thermodynamic, and Topological). Another 72 descriptors were calculated using Diverse Solutions (version 4.06; developed by R. S. Pearlman and K. M. Smith at the University of Texas, Austin, and distributed by Tripos Inc., St. Louis, MO; BCUT descriptors with explicit hydrogens).

A training set (15 000 compounds targeted, 14 431 compounds obtained from then available stock of the following vendors: ChemDiv Inc., San Diego, CA, Tripos Inc., St. Louis, MO, ChemBridge Inc., San Diego, CA, and AsInEx Inc., Moscow, Russia) was designed using the Diverse Compound Selection through a D-optimal Design strategy (Euclidian distance metric, Ochiai Similarity Coefficient, Mean/Variance Normalization, 75 000 Monte Carlo Steps at 300 K, Monte Carlo Seed of 12 379, termination after 1000 idle steps, Gaussian alpha of 1.0, bucket size of 21 for the K-d tree, and taking the nearest seven neighbors into consideration), as implemented in Cerius^2.

The training set was subsequently submitted to a proprietary high-throughput screening (HTS) procedure.

A method optimization and evaluation protocol was written that systematically varied the RP conditions implemented in Cerius^2. The terms used in this manuscript are as defined in the Cerius^2 software manual. The following conditions were considered:

*Weighting by:
- Classes
- i.e., each class is considered of equal importance to the model rather than each compound

*Splitting Method:
- Twoing
- i.e., the formalism that determines how groups are divided or partitioned into statistically distinct nodes or subgroups

*Pruning:
- Moderate
- i.e., the procedure that determines the appropriate statistically significant tree depth for each node

*Minimum Number of Samples per Node:
- 144 (1%), 90, 54, 18, 3, and 1
- i.e., a node or subgroup cannot contain fewer than this number of compounds from the training set

*Maximum Number of Knots per Split:
- systematically varied in increments of 5 starting at 5 and terminating at 200, or systematically varied using prime numbers starting at 2 and terminating at 199
- i.e., the maximum number of ways a descriptor range may be divided before statistical relevance is determined

*Maximum Tree Depth:
- 5 through 16
- i.e., the maximum number of splits that may occur before the partitioning process terminates

A collection of RP trees is generally referred to as a Forest[12,13] or an Ensemble.[14]

RESULTS AND DISCUSSION

HTS Results. The HTS procedure yielded 60 hits, which were subsequently tested in a concentration−response experiment. Thirty-seven of the 60 hits were confirmed to have significant and demonstrable activity by the concentration−response experiment, resulting in a 0.256% overall hit rate, and a 38% false positive rate. Six compounds with an EC_{50} value of ≤ 5 μM were considered "highly active" and assigned an activity class of 4. Twelve compounds with an EC_{50} value between 5 and 10 μM were considered "moderately active" and assigned an activity class of 3. Nineteen compounds with an EC_{50} value between 10 and 50 μM were considered "weakly active" and assigned an activity class

of 2. The remaining 14 395 compounds were considered "inactive" and were assigned an activity class of 1. These results represent a 0.042% hit rate for the "highly active" and most desirable compounds and a 0.125% hit rate for the "highly active" and "moderately active" compounds combined.

General Definitions. There are several measures for determining the success of an RP analysis. Cerius[2] defines and records the following for each RP tree:

1. "fold enrichment": This represents the hit rate of the model selection divided by the hit rate of the entire training set.

2. "% class correct": This is a measure 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". It is also known as "% recovery".

3. "% overall correct": This represents the total number of compounds, regardless of class, correctly classified by the model, i.e. the sum of all true positive and true negative assignments, expressed as a percentage of the entire training set.

It is relatively easy to obtain a high percent overall correct by simply classifying all compounds as inactive (99.75% overall correct) or to obtain a high percent class correct by classifying all compounds as active (100% class correct), but it is much harder to obtain a high percent class correct **and** fold enrichment while maintaining a high percent overall correct. The false positive rate (i.e. the percentage of compounds identified by the model as having a high probability of being active but not actually having demonstrable activity) and false negative rate (i.e. the percentage of compounds identified by the model as having a low probability of being active but actually having demonstrated activity) are better indicators of overall model quality. Whereas it is virtually impossible to evaluate the false negative rate of any model without experimentally testing all possible compounds repeatedly, it is feasible to evaluate the impact of model parameters on the model's false positive rate. An assay's false positive rate, however, is much more easily established (see HTS Results section).

Fold enrichment and percent class correct are not 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 are eliminated from the model. However, the method concurrently also tends to eliminate more false negatives, i.e. a higher false negative rate, resulting in a better fold enrichment in the remaining models but a lower overall percent class correct.[5]

Model Validation. A Recursion Forest of RP trees was generated using the optimization protocol. One model, our reference model, was selected from the Recursion Forest based on the criteria previously described.[5] The reference model (tree depth = 9, max. knots = 85, min. samples = 1%) predicted an 89% class correct and a 14.6-fold enrichment for the "highly active" category. By collecting all samples from terminal nodes with a class assignment of "3" or "4", 882 compounds are predicted to have an increased probability of being active. This represents a (882−18/882) or 98.2% false positive rate and a 1.816% hit rate. We maintain a database of commercially available compounds and apply certain "pharmaceutically relevant" selection criteria, such as a molecular weight cutoff of 500, a ClogP cutoff of 5, toxicity and chemical reactivity indicators, etc. Only those compounds passing all of the criteria are considered "HTS Eligible". The size of the collection is in constant flux and currently contains about 2 million compounds. We subsequently predicted against the collections of the aforementioned vendors contained in this collection which compounds had a higher probability of being active and purchased an additional set of 3417 compounds (pharmaceutically relevant exclusion criteria were also applied). These compounds were submitted to the same HTS procedure as the training set, and an additional 19 compounds were identified as "highly active", an additional 5 compounds were identified as "moderately active", and an additional 7 compounds were identified as "weakly active" (Figure 1A). These results represent a hit rate of 0.556% for the "highly active" compounds and a 0.702% hit rate for the "highly active" and "moderately active" compounds combined. The realized enrichment for this experiment is therefore 13.3-fold for the "highly active" compounds and 5.6-fold for the "highly active" and "moderately active" compounds combined (Figure 1B). The obtained fold enrichment of 13.3 is slightly lower than, but in general agreement with, the predicted fold enrichment of 14.6. Additionally, whereas fewer than 50% of the hits in the training set belong to either the "highly active" or "moderately active" categories, 77% of all hits in the validation set do.

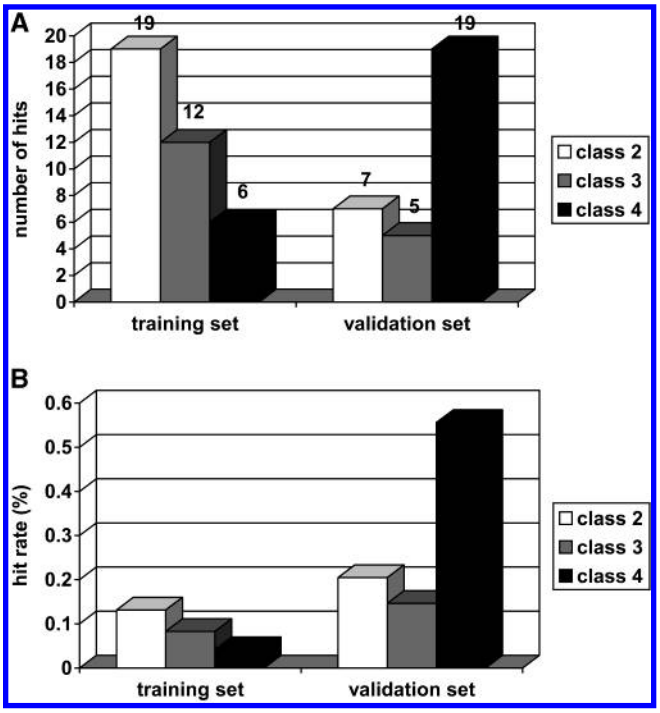

Figure 1. A: Distribution of hits in the validation set. B: Distribution of hit rates in the validation set.

Sampling Rate. The complexity of an RP tree can be thought of in the following terms:

The level of complexity of RP trees increases with increasing tree depth, with an increase in the number of knots, but decreases with larger sample size (see Table 1).

The default for the minimum number of samples in the Cerius[2] program is 1% or in our case 144 samples. Since the maximum number of "highly active", i.e. class 4, samples that we could possibly put in one terminal node is only 6,

Table 1. Variation of the "Minimum Number of Samples per Node" Criteria

tree depth	max. knots	min. samples	% class correct	fold enrichment	% false positives	no. selected	no. of active nodes	no. of terminal nodes	selected/ node
9	85	144	89	14.6	98.2	882	5	27	176.4
9	103	90	100	21.0	97.4	688	7	35	98.3
9	103	54	89	22.8	97.2	562	7	37	80.3
12	107	18	94	65.8	91.8	207	9	59	23.0
12	107	3	94	184.2	77.0	74	11	95	6.7
12	137	1	100	370.0	53.9	39	13	117	3.0

Table 2. Consensus Selection by Multiple Recursion Trees[a]

							training set					
model	descriptor	tree depth	max. knots	min. samples	% class correct	fold enrichment	compounds	total	consensus	% class correct	fold enrichment	% false positives
consensus model 1	C45	9	85	144	89	14.6	882	1245	451	78	24.8	96.9
	C45	9	90	144	83	14.8	814					
consensus model 2	C45	9	85	144	89	14.6	882	1199	632	83	19.0	97.6
	C45	9	80	144	89	13.5	949					
consensus model 3	C45	9	80	144	89	13.5	949	1493	411	73	25.3	96.8
	C45	9	85	144	89	14.6	882					
	C45	9	90	144	83	14.8	814					
consensus model 4	C45	9	85	144	89	14.6	882	996	91	83	114.3	84.6
	C45	12	107	18	94	65.8	207					
consensus model 5	C45	12	107	18	94	65.8	207	241	173	94	78.6	90.2
	C45	12	109	18	94	65.8	207					
consensus model 6	C45	12	107	18	94	65.8	207	354	77	94	176.6	77.9
	C45	12	127	18	100	64.4	224					
consensus model 7	C45	9	85	144	89	14.6	882	1563	165	83	63.0	90.9
	BCUT	8	101	144	94	15.8	848					

[a] C45 denotes Cerius[2] version 4.5 descriptors; BCUT denotes Diverse Solutions version 4.06 descriptors with explicit hydrogen atoms.

we are oversampling the training set by 24-fold, which limits the number of false positives to a minimum of 138 samples per node, i.e. at least a 95.8% false positive rate!

To split a node, $2 \times 144 = 288$ samples are required per node. In the model described above the number of samples per node varied between 145 and 237, which indicates that the RP run was likely terminated because the min. samples criteria were reached. If we lower the criteria, a larger and more complex tree can be grown, which theoretically should result in a lower false positive rate. The impact of changes to the min. samples criteria and the effect on the number of compounds actually selected per node are shown in Table 1.

Unlike the situation where model complexity increases only as a function of tree depth,[5] we found that when the number of false positives decreases as a function of the minimum node size, the percent class correct does not necessarily decrease (Table 1). However, decreasing the minimum node size does tend to slightly increase the number of knots required as well as requiring greater tree depth to achieve stability. It therefore appears that the effect of smaller minimum node size negates the effect of the greater tree depth. Consequently, a more complex model results in more terminal nodes and more active terminal nodes (see Table 1). As the false positive rate goes down, so does the number of compounds selected per node, and a more complex model also results in fewer actives per active node. In the more extreme cases the situation becomes similar to the use of the Gini method for building RP trees: high node purity biases the tree toward highly specific nodes with good explanatory power but with poor predictive power, i.e. the model can explain the training set with high accuracy, but fails readily when the model is used to predict compounds outside of the training set.[7] This is akin to overfitting in deterministic (quasi-) linear QSAR models.

Although, theoretically, we should have been able to reduce the false positive rate to zero at a "minimum number of samples per node" of 18 or less, the results as indicated in Table 1 do not bear out this possibility. As the complexity of the trees grows, so does the number of terminal nodes, and thereby the chance of undeservedly classifying compounds as active. Even at a rate of one sample per node, we obtain no less than a 53.9% false positive rate (Table 1).

When a "minimum number of samples per node" of 18 was selected, i.e. the sum of class 4 and class 3 compounds, an RP tree (tree depth = 12, max. knots = 107, min. samples = 18) could be generated predicting a 94% class correct and a 65.8-fold enrichment. This model predicted only 207 out of the original 1431 compounds selected for the model validation to have a high probability of being active. However, the model identified only three "highly active" compounds (i.e., a 1.141% hit rate) out of the original 19 present in the validation set, and an additional three "moderately active" compounds (i.e., a combined hit rate of 2.281%) out of the original seven present in the validation set (see also Table 4). This represents an actualized fold enrichment of 27.2 for the "highly active" compounds. Although substantially higher than the fold enrichment for the reference model, the model falls well short of its own predictions. Moreover, by increasing the model stringency, we have effectively eliminated 16 out of the 19 originally identified "highly active" compounds, i.e. a false negative rate of 84.2%. What the experiment described here does not address, however, is how many "highly active" compounds

Table 3. Consensus Selection Applied to HTS Eligible Compounds[a]

							HTS eligible			
model	descriptor	tree depth	max. knots	min. samples	% class correct	fold enrichment	compounds (thousands)	total no. (thousands)	consensus no. (thousands)	concentration factor
consensus model 1	C45	9	85	144	89	14.6	73	156	26	5.9
	C45	9	90	144	83	14.8	69			
consensus model 2	C45	9	85	144	89	14.6	73	168	50	3.3
	C45	9	80	144	89	13.5	106			
consensus model 3	C45	9	80	144	89	13.5	106	235	22	10.8
	C45	9	85	144	89	14.6	73			
	C45	9	90	144	83	14.8	69			
consensus model 4	C45	9	85	144	89	14.6	73	95	6	15.2
	C45	12	107	18	94	65.8	28			
consensus model 5	C45	12	107	18	94	65.8	28	47	20	2.4
	C45	12	109	18	94	65.8	26			
consensus model 6	C45	12	107	18	94	65.8	28	61	10	6.1
	C45	12	127	18	100	64.4	30			
consensus model 7	C45	9	85	144	89	14.6	73	128	6	20.3
	BCUT	8	101	144	94	15.8	61			

[a] C45 denotes Cerius[2] version 4.5 descriptors; BCUT denotes Diverse Solutions version 4.06 descriptors with explicit hydrogen atoms.

could have been identified if the model were applied to the entire collection, rather than just the validation set. In effect, this experiment only describes the Consensus Selection between the two models.

Consensus Selection. Consensus Selection is a process for group decision-making. It is a method by which a group of models can be in agreement. The input and statistics of all participating models are gathered and synthesized to arrive at a final model satisfying the conditions of all contributing models. Voting (aka election) is a means by which one model is preferentially selected from several models by weighing the input of each of the individual models. Consensus Selection, on the other hand, is a process of synthesizing many diverse elements together.

The process involves the determination of the Boolean intersection of a set of models (at least 2, in theory unlimited, individually derived models), thereby emphasizing the probabilities of the consensus set and de-emphasizing the probabilities of the contributors for each of the models excluded from the consensus set, i.e. the dissenting sets. The process is expected to have a higher chance of eliminating false positives from the process, thereby reducing operating cost and throughput requirements, shortening timelines, and increasing the reliability of the process.

The Consensus Selection methodology has been used previously in alignment problems, e.g., refs 15 and 16, and in deterministic modeling procedures, e.g., ref 17, but has, until now, not been associated with probabilistic modeling methods such as RP.

Table 2 describes various ways to derive models using Consensus Selection by RP trees.

Whereas the maxim known as "Occam's Razor" or "Ockham's Razor" would lead us to select the single most parsimonious hypothesis from among multiple hypotheses proposed, Consensus Selection directs us to synthesize a new hypothesis from its predecessors. This is especially useful when Ockham's Razor is hard to apply, such as in situations where near-identical models yield nearly indistinguishable results. The simplest solution, in this case, is to not select a single hypothesis but to combine useful elements from all contributing hypotheses.

Table 2 describes the results if models of similar complexity are paired (consensus models 1, 2, and 5) or grouped (consensus model 3) together. Table 2 also describes the results when models are not entirely equivalent (consensus model 6) or purposely mismatched by complexity (consensus model 4) or descriptor basis (consensus model 7).

Consensus model 1 describes the Boolean intersection of the reference model and a slightly more complex model. As can be seen in Table 2, similar models behave similarly with respect to percent class correct and fold enrichment. However, when Consensus Selection is applied, the number of compounds selected drops from 882 (tree depth = 9, max. knots = 85, min. samples = 144) or 814 (tree depth = 9, max. knots = 90, min. samples = 144) to 451, which is almost a 50% reduction in the total number of compounds selected but translates into only a relatively small change in the false positive rate. A 50% decrease in the number of compounds selected without loss of positives would double the fold enrichment of the process. In theory, the percent class correct for the consensus model can be no higher than the percent class correct of the worst performing model in the set. Consensus model 1 demonstrates that the percent class correct can be negatively affected by employing Consensus Selection. It is therefore important to evaluate how closely the individual models are matched.

Consensus model 2 describes the Boolean intersection of the reference model and a slightly less complex model. The less complex model itself does not meet the selection criteria outlined earlier[5] as it is closer (too close) to an instable region in the model optimization trace. In this case, the models match their respective percent class correct but have different outcomes for fold enrichment and the number of compounds predicted to have an increased probability of being active. Whereas the percent class correct for consensus model 2 (83%) is higher than that for consensus model 1 (78%) (see Table 2), the number of compounds selected is only reduced by 30% (Table 3) and without apparent effect in the validation set (Table 4).

Consensus model 3 describes the Boolean intersection of the reference model and both models of lesser and higher complexity. It is therefore expected to have a percent class correct of no better than the worst performing contributing model (83%) and a fold enrichment no worse than the best performing contributing model (14.8). Indeed, consensus model 3 has a 73% class correct and prioritizes only 411

Table 4. Consensus Selection Applied to the Validation Set[a]

consensus model	model parameters: descriptor	tree depth	max. knots	min. samples	percent class correct	fold enrichment	compds	total	consensus	class 4 hits	class 3 hits	class 2 hits	validation set: percent class 4 correct	percent class 3 correct	percent class 2 correct	fold enrichment class 4 only	percent class correct 4,3 combined	percent class correct 4,3,2 combined	fold enrichment 4,3 combined	fold enrichment 4,3,2 combined
# 1	C45	9	85	144	89	14.6	2242	2323	1351	18	4	5	95	80	71	31.7	92	87	13.0	7.8
	C45	9	90	144	83	14.8	1432													
# 2	C45	9	85	144	89	14.6	2242	2342	1553	18	5	5	95	100	71	27.6	96	90	11.8	7.0
	C45	9	80	144	89	13.5	1653													
	C45	9	80	144	89	13.5	1653													
# 3	C45	9	85	144	89	14.6	2242	2395	1270	18	4	5	95	80	71	33.7	92	87	13.9	8.3
	C45	9	90	144	83	14.8	1432													
# 4	C45	9	85	144	89	14.6	2242	2267	238	3	3	1	16	60	14	30.0	25	23	20.2	11.5
	C45	12	107	18	94	65.8	263													
# 5	C45	12	107	18	94	65.8	263	314	210	2	3	1	11	60	14	22.7	21	19	19.0	11.1
	C45	12	109	18	94	65.8	261													
# 6	C45	12	107	18	94	65.8	263	474	92	2	1	1	11	20	14	51.8	13	13	26.1	17.0
	C45	12	127	18	100	64.4	303													
# 7	C45	9	85	144	89	14.6	2242	2365	421	9	4	4	47	80	57	50.9	54	54	24.7	2.8
	BCUT	8	101	144	94	15.8	544													

[a] C45 denotes Cerius[2] version 4.5 descriptors; BCUT denotes Diverse Solutions version 4.06 descriptors with explicit hydrogen atoms. Note: Of the 3417 compounds originally available from our vendors and included in the validation set, only 2242 are still available at the time of this writing. The numbers in the table reflect this.

compounds (Table 2), a 70% reduction in projected test set size (Table 3). It is anticipated that including more models in the Consensus Selection process will further reduce the percent class correct. How many models can be practically included in the process is therefore more an economical than a scientific question. Although, this leaves unresolved how closely RP trees need to be matched before Consensus Selection can be applied.

We have previously observed that starting with a default setting of 20 for the "maximum number of knots per split" of the RP procedure as implemented in Cerius[2], and incrementing the value in steps of 5, can lead to a certain periodicity in the optimization traces.[5] This would indicate that there is an inter-relationship between such models that overrides or coincides with the splitting criteria used to obtain statistically significant splits. We have since changed to a procedure using prime numbers as the max. knots setting and have not seen similar periodicity in the optimization traces (results not shown). Use of prime numbers, however, limits the number of possible models within a stable region of the optimization traces and restricts the coarseness of the internal similarity of the RP trees, since they occur at irregular and unevenly spaced intervals. So far, we have been unable to resolve this duality between the apparent need for closely matched RP trees as contributors to Consensus Selection and the constraints imposed by the RP tree building and optimization methods.

All three consensus models described above compare favorably to the individual contributing models when compared by total number of compounds prioritized. In this particular case, the number of correctly identified highly active compounds is identical for all three preceding Consensus Selections (Table 4). However, a small decrease in the number of correctly identified compounds in classes 3 and 2 can be observed (Figure 1A and Table 4).

To determine how closely related the various models need to be for an effective Consensus Selection process, we investigated the Boolean intersection of two models that satisfy the selection criteria of their individual optimization traces. The first model (tree depth = 9, max. knots = 85, min. samples = 144), our reference model, is far less complex than the second model (tree depth = 12, max. knots = 107, min. samples = 18) (see Table 1). With a high percent class correct, it is not expected that the more complex model would interfere with the efficiency of the less complex model. Indeed, consensus model 4 shows a considerable decrease in the false positive rate (Table 2) but at the same time is only marginally better than consensus model 1, and no better than consensus model 2, with regard to percent class correct. However, the reduction in the number of compounds prioritized is substantial: up to 91% based on the less complex model and up to 78% based on the more complex model (Table 3). Conversely, when the consensus model was applied to the validation set, only three out of 19 class 4 compounds (i.e. a false negative rate of 84.2%) and an additional four out of 12 class 3 or class 2 compounds could be accurately identified. It appears that the more complex model suffers from overfitting, which causes very few actives to be classified with high accuracy, and limits the general applicability of the model. The application of Consensus Selection procedures clearly cannot overcome the imbalance introduced by mixing dissimilar models. Use of

Consensus Selection methods therefore appears to be restricted to contributing models that are similar not only in their output performance characteristics but also in their internal complexity.

Consensus model 5 demonstrates that higher efficiencies can be obtained by using Consensus Selection on higher complexity models (Table 2). A 94% class correct and a 90.2% false positive rate could be obtained by selecting two similar models of higher complexity than the reference model (tree depth = 12, max. knots = 107, min. samples = 18, and tree depth = 12, max. knots = 109, min. samples = 18, respectively). However, the gain in efficiency can be wiped out when the projections into the HTS eligible compound collection are taken into account (Table 3). The set of compounds prioritized by consensus model 5, at 19 720 compounds, is only nominally smaller than the 21 821 compounds prioritized by consensus model 3. This tentatively reflects an uneven chemical diversity distribution in the HTS eligible compound collection. Notably, the number of selected compounds exceeds the number of compounds in the training set, and the model therefore does not represent a gain in overall screening efficiency, rather it indicates a lack of selectivity in the compound selection process. It therefore appears that some degree of divergence is needed between individual RP trees to economically employ Consensus Selection in the sequential screening process.

Consensus model 6 was created to study the impact of selecting slightly dissimilar contributing models. The second contributing model (max. knots = 127) is only marginally more complex than the reference model (max. knots = 107) but exhibits an exceptionally high percent class correct: 100%. As shown in Table 2, a high percent class correct is retained in the consensus model and a remarkable reduction in the false positive rate can be achieved. Table 3 indicates that a reduction of as much as 67% of the number of prioritized compounds, which boosts the theoretical fold enrichment to about 180-fold, can be obtained under favorable circumstances.

However, the reference model is not part of consensus models 5 and 6 and falls outside the scope of the loosely defined "similar models" criteria above. Consequently, the data in Table 4 cannot be used directly to support or to refute the hypotheses and is an unintended and unanticipated consequence of the experimental design.

The final consensus model, consensus model 7, was created to investigate the contribution of the descriptor base to the Consensus Selection process. The reference model (tree depth = 9, max. knots = 85, min. samples = 144) was created using the descriptor base available in Cerius[2] (version 4.5), and the alternate model (tree depth = 8, max. knots = 101, min. samples = 144) was created using the descriptor base available through DiverseSolutions (version 4.0.6). In theory, it would be preferable to derive contributor models from independent descriptor bases, since this would eliminate bias introduced by, e.g., systematic error or descriptor type selection by a vendor, a programmer, or the optimization algorithm. In this example, we combined two independently derived and optimized models of similar complexity to address this. As is evident from Table 2, the contributing models behave very similarly at the gross performance level, such as percent class correct (89 and 94, respectively), fold enrichment (14.6 and 15.8, respectively), or number of compounds prioritized (882 and 848, respectively), and they are relatively similar in terms of their internal complexity. The consensus model still classifies 15 out of the 18 most active compounds correctly, i.e. 83% class correct, whereas the false positive rate has decreased considerably to 90.9% (Table 2). The projection of potential utility into the HTS eligible compound collection is much better than consensus model 1, 2, or 3 of comparable complexity and at least as good as consensus model 6 of much greater complexity (Table 3). However, validation of the model (Table 4) indicates that pairing independent descriptor bases readily results in loss of a significant number of active compounds, resulting in a false negative rate of 52.6% for class 4 only and a false negative rate of 45.8% for class 4 and class 3 combined. This result was not anticipated and is not readily explained. We speculate, however, that this is the result of an inadequate or incomplete parametrization of chemical diversity space by the individual descriptor bases. An investigation on the impact of descriptor base selection will be reported separately.

The above observations lead us to hypothesize that (1) a high percent class correct and (2) small but significant divergence between RP trees are required to effectively leverage Consensus Selection. However, we cannot verify these assertions and projections because of the previously noted problems with the current validation set. This warrants further investigation.

CONCLUSIONS

We have demonstrated that RP can be used to augment the sequential screening process.[5] Here we show that RP suffers from a high false positive rate and that corrections can be introduced to the RP Forest building and optimization problem that can ameliorate the process. Whereas strict reductions in the "minimum number of samples per node" can attenuate the false positive rate, and while retaining good explanatory power, the RP trees seem to lose their capability to predict outside of the training set. Conversely, use of Consensus Selection by Multiple RP Trees has poor explanatory power, because only the probabilities of the contributing RP trees are considered, but the method has good predictive power. We have provided experimental evidence that Consensus Selection by Multiple Recursion Trees is superior to the use of single RP trees when applied in the sequential screening process. We have shown that in excess of 30-fold enrichment can be obtained using this method and that better than 70% class correct can be retained while significantly reducing the false positive rate. We fully expect these improvements to the process to further reduce the occurrence of false positives from the process, thereby reducing operating cost and throughput requirements, shortening timelines, and increasing the reliability of the process.

ACKNOWLEDGMENT

I thank Drs. Kerry Spear and Brian Marron and Ms. Laura Van Zant for helpful discussions and editorial assistance during the preparation of this manuscript and the reviewers for their thoughtful consideration and suggestions.

REFERENCES AND NOTES

(1) Abt, M.; Lim, Y. B.; Sacks, J.; Xie, M.,; Young, S. S. A Sequential Approach for Identifying Lead Compounds in Large Chemical Databases. *Natl. Inst. Stat. Sci., Technical Report* **2002**, *105*, 1−32.

(2) Holzgrabe, U.; Mohr, K. Allosteric Modulators of Ligand Binding to Muscarinic Acetylcholine Receptors. *Drug Discovery Today* **1998**, *5*, 214−222.
(3) Zwart, R.; Vijverberg, H. P. Potentiation and Inhibition of Neuronal Nicotinic Receptors by Atropine: Competitive and Noncompetitive Effects. *Mol. Pharmacol.* **1997**, *52*, 886−895.
(4) 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.
(5) van Rhee, A. M.; Stocker, J.; Printzenhoff, D.; Creech, C.; Wagoner, P. K.; Spear, K. L. Retrospective Analysis of an Experimental High-Throughput Screening Data Set by Recursive Partitioning. *J. Combi. Chem.* **2001**, *3*, 267−277.
(6) Rusinko, A.; Young, S. S.; Drewry, D. H.; Gerritz, S. W. Optimization of Focused Chemical Libraries Using Recursive Partitioning. *Comb. Chem. High Throughput Screening* **2002**, *5*, 125−133.
(7) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees*; Wadsworth: Boca Raton, FL, 1984.
(8) 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.
(9) Young, S. S.; Hawkins, D. M. Analysis of a 2^9 Full Factorial Chemical Library. *J. Med. Chem.* **1995**, *38*, 2784−2788.
(10) 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.
(11) Harper, G.; Bradshaw, J.; Gittins, J. C.; Green, D. V. S.; Leach, A. R. Prediction of Biological Activity for High-Througput Screening Using Binary Kernel Discrimination. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1295−1300.
(12) Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5−32.
(13) Ho, T. K. The Random Subspace Method for Constructing Decision Forests. *IEEE Trans. Pattern Analysis Machine Intelligence* **1998**, *20*(8), 832−844.
(14) Dietterich, T. An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. *Machine Learning* **2000**, *40*(2), 139−157.
(15) Ravi, M.; Hopfinger, A. J.; Hormann, R. E.; Dinan, L. 4D-QSAR Analysis of a Set of Ecdysteroids and a Comparison to CoMFA Modeling. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1587−1604.
(16) 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.
(17) Kastenholz, M. A.; Pastor, M.; Cruciani, G.; Haaksma, E. E. J.; Fox, T. GRID/CPCA: A New Computational Tool to Design Selective Ligands. *J. Med. Chem.* **2000**, *43*, 3033−3044.

CI034023J