Enrichment of High-Throughput Screening Data with Increasing Levels of Noise Using Support Vector Machines, Recursive Partitioning, and Laplacian-Modified Naive Bayesian Classifiers

Meir Glick,*,[†] Jeremy L. Jenkins,[†] James H. Nettles,[†] Hamilton Hitchings,[‡] and John W. Davies[†]

Lead Discovery Center, Novartis Institutes for Biomedical Research Inc., 250 Massachusetts Avenue,
Cambridge, Massachusetts 02139, and Equbits LLC, 2625 Middlefield Road, #102,
Palo Alto, California 94306

Received September 4, 2005

High-throughput screening (HTS) plays a pivotal role in lead discovery for the pharmaceutical industry. In tandem, cheminformatics approaches are employed to increase the probability of the identification of novel biologically active compounds by mining the HTS data. HTS data is notoriously noisy, and therefore, the selection of the optimal data mining method is important for the success of such an analysis. Here, we describe a retrospective analysis of four HTS data sets using three mining approaches: Laplacian-modified naive Bayes, recursive partitioning, and support vector machine (SVM) classifiers with increasing stochastic noise in the form of false positives and false negatives. All three of the data mining methods at hand tolerated increasing levels of false positives even when the ratio of misclassified compounds to true active compounds was 5:1 in the training set. False negatives in the ratio of 1:1 were tolerated as well. SVM outperformed the other two methods in capturing active compounds and scaffolds in the top 1%. A Murcko scaffold analysis could explain the differences in enrichments among the four data sets. This study demonstrates that data mining methods can add a true value to the screen even when the data is contaminated with a high level of stochastic noise.

INTRODUCTION

High-throughput screening (HTS) is commonly used in lead discovery efforts. The immense amount of data which is being generated provides a large resource for data mining using cheminformatics approaches. One example is an iterative screen or a "sequential approach".[1] In an attempt to reduce the time required to screen a large compound collection, deal with reagent and disposal costs, and reduce compound collection depletion,[2] a focused or diverse subset of the compound collection can be screened instead of the full deck.[3] Cheminformatics approaches are then utilized to expand the number of compounds and chemotypes by building a statistical model based on the screened subset and utilizing this model in order to design the next batch of compounds to be screened. Data mining is also employed to prioritize hit lists when screening in mixtures[4] where more than one compound is pooled into the same well. The deconvolution of the primary data where each compound in the mixture has to be tested as a single compound is time-consuming. Glick and co-workers[4] showed that a statistical model built from the nondeconvoluted primary data, that is, the mixture, could prioritize the "hits" and identify the true positives. Screening single compounds also generates noisy data, particularly in cell-based or miniaturized assay formats, for example, in 384 and 1536 well plates. Therefore, there is an obvious motivation to devise data mining methods that are tolerant to noise in order to triage the "hits"—identify false positives and negatives, maximize the number of chemotypes, and build structure activity relationships from the primary screening data. If we ignore, for the moment, the proper selection of chemical descriptors[5] and focus on the nature of HTS data, devising an accurate and reliable model based on such data is a difficult task.

First, the definition of an active compound is not absolute. Primary HTS normally measures efficacy (% inhibition) and not potency (IC_{50} or EC_{50}). A cutoff of 50% inhibition means that a compound with 50.1% inhibition is "active" and one with 49.9% inhibition is "inactive". However, in practical terms, the latter compound might turn out to be more potent than the former one. Sills and co-workers[6] demonstrated in three assay formats, AlphaScreen, time-resolved fluorescence (TRF), and time-resolved fluorescence resonance energy transfer (TR−FRET), that the common assumption that the same hits would be found regardless of which assay technology is being used is not valid. Even for the same experiment, different data normalization algorithms can be employed.[7]

Second, HTS data sets are unbalanced. In a typical HTS campaign, the hit rate is 0.1−1%. The active compounds will, therefore, have a low occurrence and be a minority class. One should, therefore, ensure that the minority class will be preserved for the model building; that is, the subsets chosen for the model building should not be strictly random but have a significant number of observations of the low-occurrence class (the "active" compounds). In other words, the relative importance of the minority class should be taken into

* Corresponding author e-mail: meir.glick@novartis.com; phone: 617-871-7130.
[†] Novartis Institutes for Biomedical Research Inc.
[‡] Equbits LLC.

10.1021/ci050374h CCC: $33.50
Published on Web 12/03/2005

consideration. In this study, we reduced the number of "inactive compounds" in the training set by employing a diversity selection (vide infra). Therefore, in the training set, we increased the "active/inactive compounds" ratio. Although this ratio has been enlarged, the number of "actives" was still significantly lower than the number of "inactives".

Third, a HTS experiment is noisy in terms of false positives and negatives that reduce the signal-to-noise ratio.[8] Assay artifacts are inherent to the chemical structure being screened (reactive, fluorescent, quencher, aggregator, chelator, cytotoxic, and reducing compounds), the assay (reagent and temperature errors), screening technology (liquid handling, reader, and time errors), and the storage conditions (compound precipitation from DMSO stock solutions). Given the low hit rate in a HTS campaign, we postulate that false negatives are simply a loss of information, while the primary source of noise is the false positives. False positives might be due to stochastic (random) or systematic (such as frequent hitters) errors. Both types will skew the accuracy of the model. However, many systematic errors, such as compounds that are frequent hitters in a certain assay format, can be identified and removed before the model is generated.

Fourth, unlike a lead optimization project where one or a few chemotypes, from congeneric series that demonstrate activity and selectivity in a biological assay, are being optimized in attempt to identify a clinical candidate, a screening campaign can consist of tens of thousands of different chemotypes. Sparse 2D descriptors can somewhat address the vast chemical diversity of the data sets[1]. As a result, the size of the data matrix (number of compounds times the number of descriptors) needed to describe an HTS campaign can be immense. The large number of descriptors raise the problem known as the "curse of dimensionality".[9] The convergence of any estimator to the true value of a smooth function defined on a space of high dimension is very slow. The number of samples (compounds) per variable (descriptor) increases exponentially with the number of descriptors to maintain a given level of accuracy. Therefore, data mining methods which scale linearly with respect to the number of compounds or features are considered necessary.

Simple selection strategies such as nearest-neighbor and substructure searches may be employed to suggest putative active compounds. However, given the noisy nature of primary HTS data, this paper will attempt to judge the robustness of three machine learning methods—Laplacian-modified naive Bayesian,[10] recursive partitioning[1] (RP), and support vector machine[11] (SVM) classifiers—with various levels of stochastic noise. The Laplacian-modified naive Bayesian classifier was successfully utilized to enrich noisy HTS data when screening in mixtures[4] and high-throughput docking data[12,13] and to classify kinase inhibitors.[10] RP was employed to enrich mixtures' data, identify multiple mechanisms of action, and generate structure−activity relationships from very large structure−activity data sets[1,14−16]. SVMs were recently applied to the drug discovery field in drug-likeness prediction and for the identification of additional active compounds in few data sets.[17]

METHODS

Enrichment and ROC curves. Various metrics such as receiver operating characteristic (ROC) curves are available to measure the robustness of the model in classifying data samples.[18] Enrichment and ROC curves are closely related, and attempting to answer the question of ranking and testing compounds in a particular order is beneficial. To generate these curves, the classifier is employed to rank the compounds in a decreasing activity probability. An enrichment curve is a cumulative count of the number of "active" compounds captured when testing them in such order. A ROC curve describes the tradeoff between sensitivity and specificity. Sensitivity is defined as the ability of the model to avoid "false negatives", while specificity is its ability to avoid "false positives". In practical terms, the "false positives" rate is plotted on the x axis and the "true positives" rate on the y axis. An ideally sensitive classifier would have an area under the curve of 1, while a random classifier will have a value of 0.5. Roughly, an area between 0.7 and 0.8 is considered as reasonable, 0.8−0.9 as good, and 0.9−1 as excellent.[12]

Laplacian-Modified Naive Bayesian Classifier. The Laplacian-modified naive Bayesian classifier has been described in detail by Xia and co-workers.[10] In principle, the Laplacian-corrected estimator for a compound being active given a feature F_i, $P_{\mathrm{corr}}(\mathrm{Active}|F_i)$, is calculated according to the following equation: $P_{\mathrm{corr}}(\mathrm{Active}|F_i) = [A + P(\mathrm{Active})\ K]/(B + K)$, where a feature F_i is contained in B samples and A of those B samples are active. An estimate of the baseline probability of a randomly chosen compound being active, $P(\mathrm{Active})$, is simply the ratio between the active compounds and the total number of compounds selected for the training. K is a constant used to add virtual samples of $P(\mathrm{active})$ in order to stabilize the estimator when the value of B is too low. Once the values of $P_{\mathrm{corr}}(\mathrm{Active}|F_i)$ are known, one can calculate a relative estimate, $P_{\mathrm{final}}(\mathrm{Active}|F_i)$, in the following manner: $P_{\mathrm{final}}(\mathrm{Active}|F_i) = P_{\mathrm{corr}}(\mathrm{Active}|F_i)/P(\mathrm{Active})$.

More than one feature is normally required in order to characterize a compound. Therefore, the multiple features F_i in a sample have to be combined. Given n features for a compound, the combined estimation, P_{combined}, is calculated as follows: $P_{\mathrm{combined}} = P_{\mathrm{final}}(\mathrm{Active}|F_1) \times P_{\mathrm{final}}(\mathrm{Active}|F_2)... \times P_{\mathrm{final}}(\mathrm{Active}|F_n)$. The Scitegic[19] implementation in PipelinePilot 4.1 was employed in this work.

Recursive Partitioning. In RP, the group of compounds is recursively partitioned into two or more statistically distinct nodes until the response (activity of a compound) variable is homogeneous. At the end of the process, the compounds are classified into similar response nodes. Splitting criteria of Gini and χ-squared were used to measure the node impurity and to determine where to make the split. In the case of the χ-squared criterion, a χ-squared test is applied between the two putative child nodes in order to identify the best possible split that yields statistically different child nodes. The Gini index is calculated as follows: if a node has a proportion of p_j of each of the classes (descriptors), then the Gini index is $1 - \sum_j P_j^2$. The algorithm normally selects the split that most decreases the Gini index.

Bootstrapping is based on the procedure of sampling with replacement; the same sample can be picked more than once. The repeats are then removed, and therefore, a certain percent of the unique samples from the original data set appear in the final sample. The process of growing multiple trees from bootstrap samples of the data is called bagging. Combining the results from multiple trees, for example, by taking the average of the predictions from all the trees, can provide a more stable prediction than a single tree.[9,18]

Various implementations of RP exist, such as SCAM[1], CART,[20] C4.5,[21] FACT,[22] and LMDT.[23] This paper is far from a complete reference for the field of classification trees. Hastie et al.[9] have provided an excellent additional resource on the theory behind classification trees. In this work, we used SciTegic's Decision Trees Collection (DTC) in PipelinePilot 4.1.[19] The DTC relies on NovoD ArborPharm, developed by NovoDynamics, Inc.[24] The main reason for choosing the above implementation is its ability to model the amounts of data in the HTS data sets.

Support Vector Machines. SVM methods, which were originally developed by Vapnik,[11] can be used to derive a hyperplane that separates the active from the inactive compounds. This maximum margin hyperplane separates the two classes of compounds by the maximum distance possible in the feature space. It is done by selecting a small subset of the critical instances from the active and inactive compounds called support vectors and using them build a linear discriminant function, which separates the active and inactive compounds as widely as possible. It is possible and practical to include nonlinear terms by transforming the instance space into a new space. A line taken from the new space will not look straight in the original instance space.[18]Various kernel functions can be used to implement different instance space transformations. The radial basis function was used as the kernel function in this work.[9] SVM can also be expanded into cases where the training data is not separable by a line either in the instance space or in the new transformed space, by allowing a few of the instances to be on the wrong side of the margin and defining slack variables. It is up to the user to tune the tradeoff between the size of the hyperplane margin and the number of correctly classified instances near the decision boundary. The optimal value for the tuning parameter, γ, was estimated by 10-fold cross validations where the weighting between precision and recall was equal. An initial value of 0.03 was used for γ. This value is based on previous experience with this type of data, followed by an automatic local optimization by the Equbits Foresight SVM method.[25] Foresight was able to apply SVMs successfully to these data sets because of its scalability and automated model tuning.

Molecular Descriptors. A plethora of scientific papers attempt to compare the effectiveness of structural descriptors.[26−29] Recently, Hert and co-workers[30] conducted an exhaustive comparison of topological descriptors for similarity-based virtual screening on data sets taken from the MDL Drug Data Report. Four types of descriptors have been reviewed: structural keys, hashed fingerprints, circular substructure descriptors, and pharmacophore vectors. Extended connectivity fingerprints (ECFPs) were by far the most effective 2D molecular representation considered and were, therefore, selected for this study. The derivation of the ECFPs was described elsewhere.[13,30] ECFPs rely on the Morgan algorithm.[31] In principle, the algorithm assigns an initial code to each atom. Then, each atom code is updated in an iterative manner to reflect the codes of each atom's neighbors. When the desired neighborhood size is reached, the process is complete and the set of all features are returned. The final fingerprint is the collection of all features generated for each atom at each iteration level; that is, a hashing function then generates the new ECFP code from the codes of an atom and its neighbors. ECFPs contain a significant number of different structural units that are crucial for the molecular comparison among the compounds. We assume that the large set of features makes ECFPs suitable for the analysis of the chemically diverse HTS data. We employed a neighborhood size of six for all the test cases in this study.

The Laplacian-modified naive Bayesian classifier and the recursive partitioning algorithms we employed could handle the high dimensionality of the ECFPs. The SVM algorithm used two approaches to handle the high dimensionality of the ECFPs. The first one used by SVM was a folded fingerprint approach whose results are presented in detail in this paper. The ECFPs codes were folded into 2048 bits according to the following procedure. For each compound, 2048 bits were allocated with an initial value of 0. The algorithm then looped over the ECFP codes of each atom for a given compound and divided each of the ECFP codes by 2048 and kept the absolute remainder values (ECFP codes are integers). Bits were set from 0 to 1 if their index (i.e., location in the 2048-bit array) was identical to the absolute remainder values from the division. This process was repeated for each of the compounds. The second approach used by SVM was the nonfolded fingerprint one, which performed feature reduction of the ECFPs in order to retain the top fingerprints that contained the most information. Preliminary results demonstrated improved results over the first approach, for the area under the ROC curve and reducing the SVM model complexity. In addition, this approach allowed the ECFPs with the greatest contribution to be retained for further analysis.

Scaffold Definition. From the medicinal chemistry point of view, the diversity of a given data set is correlated to its number of different atomic frameworks. Here, we employed the Bemis and Murcko definition[33] where the aliphatic side chains are trimmed. The rings' and the linker atoms' types, hybridizations, and bond orders remain untouched. This atomic framework will be referred to as the "Murcko scaffold".

Diversity Selection. The diversity selection was done using the "diverse molecules" component and ECFPs in Pipeline Pilot, where the selection is based on a maximum dissimilarity approach. It begins by randomly choosing a compound. The compound maximally distant from the first chosen compound is selected as the next compound. The compound maximally distant from both compounds is the next one, and so on.

RESULTS

The Data Sets. Four recent HTS campaigns where ca. 650 000 compounds were tested for their inhibitory activity were selected for this study. These compounds were either purchased from external vendors based on their drug likeness[32] or were synthesized in-house for lead optimization projects and exploratory chemistry. The screening campaigns were employed on three target classes—protease, chemokine receptor, and G-protein-coupled receptor—using either a cell-based or a biochemical assay (Table 1). In primary screening, compounds were tested as one compound per well at a concentration of 10 μM. The average Z' factors for these

Table 1. HTS Data Sets Selected for This Study

data set #	target class	assay type		# of active compounds	# of inactive compounds	total # of active and inactive compounds
1	protease	cellular (reporter gene)	training	228	4200	4428
			test	118	171 608	171 726
2	chemokine receptor	biochemical (receptor−ligand)	training	94	4197	4291
			test	46	171 730	171 776
3	G-protein-coupled receptor	biochemical (receptor−ligand)	training	318	4197	4515
			test	148	171 434	171 582
4	chemokine receptor	biochemical (receptor−ligand)	training	258	4199	4457
			test	128	171 432	171 560

screens were higher than 0.8. Normally, compounds of at least 50% inhibition were considered as "hits" and were then retested in dose response curves (IC_{50}). The percent inhibition is the difference between the signal of the sample well and the median of the signals obtained from the blank wells on the same plate divided by the difference between the median signal of the control wells (wells with the target protein but without the test compound) and the median signal of the blank wells that are located on the same plate. A compound confirmed with an $IC_{50} < 20\ \mu M$ was defined as "active" in this study. Each of the data sets was divided into training and test sets as shown in Table 1. The active compounds were divided at random between the training and test sets. The 171 560−171 776 inactive compounds for the test sets were selected at random from the inactive subset of the screen. The 4197−4200 inactive compounds for the training sets were based on a diversity selection from a different subset of inactive compounds. Apart from the speedup in training the classifier (less-inactive compounds are needed to train the classifier), diversity and not random selection of inactive compounds normally produces more accurate models (data not shown). It is perhaps due the fact that the diversity of the screening collection is not uniform since congeneric series of compounds from lead optimization projects are often being added to the screening deck. Infrequent chemical classes that may exemplify a different subset in the "chemical space" might not be picked up at all if a random selection is being employed. It should be noted that data set 1 resulted from a cell-based assay, while data sets 2−4 were biochemical assays.

The Enrichment and Area under the ROC Curve for the Four Data Sets. Before building any model in this work, it should be noted that systematic errors (such as frequent hitters) and artifactual compounds (such as compounds that are not stable in solution) were removed. Six statistical models were built from each of the training sets as shown in Figure 1a and b. Three models included recursive partitioning based on 10 or 100 trees and χ-squared or Gini as the splitting criteria. To understand the impact of folding the fingerprint (or loss of information) on the accuracy of the model, two Laplacian-modified naive Bayesian models were generated. In the first model, the fingerprint was folded into 2048 bits, and in the second, the fingerprint remained unfolded. And lastly, support vector machines model was constructed using the 2048-bit folding scheme. Both in terms of the percentage of active compounds captured in the top 1% (Figure 1a) and the area under the ROC curve (Figure 1b), data set 1 seemed to be the most difficult to enrich. Unlike data sets 2−4, where the percentage of actives captured in the first 1% ranged from 26% (data set 3, RP, 10 trees, χ squared) to 55.7% (data set 4, 2048-bit folding scheme, SVM), the values for data set 1 ranged between 12.7% and 22%. The maximal value of the area under the ROC curve for data set 1 was 0.79 (Laplacian-modified naive Bayes), which did not reach even the minimal value of the other three data sets (data set 2, 2048-bit folding scheme, SVM). In terms of the percent of actives captured in the first 1%, SVM with the 2048-bit folding scheme outperformed any of the other classifiers. In terms of the area under the ROC curve, the Laplacian-modified naive Bayesian was shown to be the most robust method in data set 1. For data sets 2−4, the differences between the classifiers were not prominent.

The Impact of False Positives and False Negatives on the Accuracy of the Models. For each of the data sets, we generated artificial stochastic noise of false positives by picking compounds at random from the inactive subset in the training set and changing their attribute to "active". Two levels of noise were introduced for each of the four data sets: a ratio of 1:1 between the true active compounds and the misclassified compounds and a ratio of five misclassified compounds for each active compound in the training set (noise ratio of 5:1). For each data set and level of noise, three classifiers were generated: recursive partitioning based

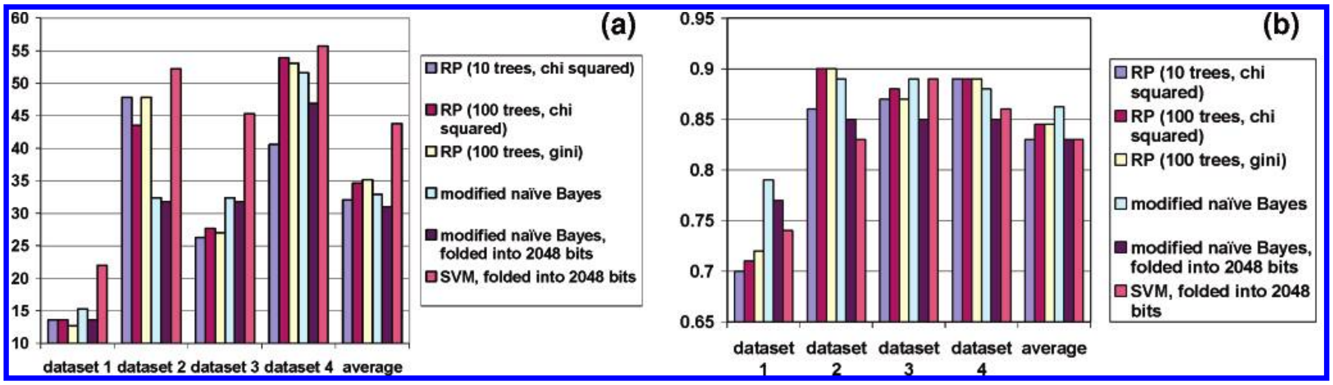

Figure 1. (a) Percentage of active compounds captured in the top 1%. (b) Area under the ROC curve.

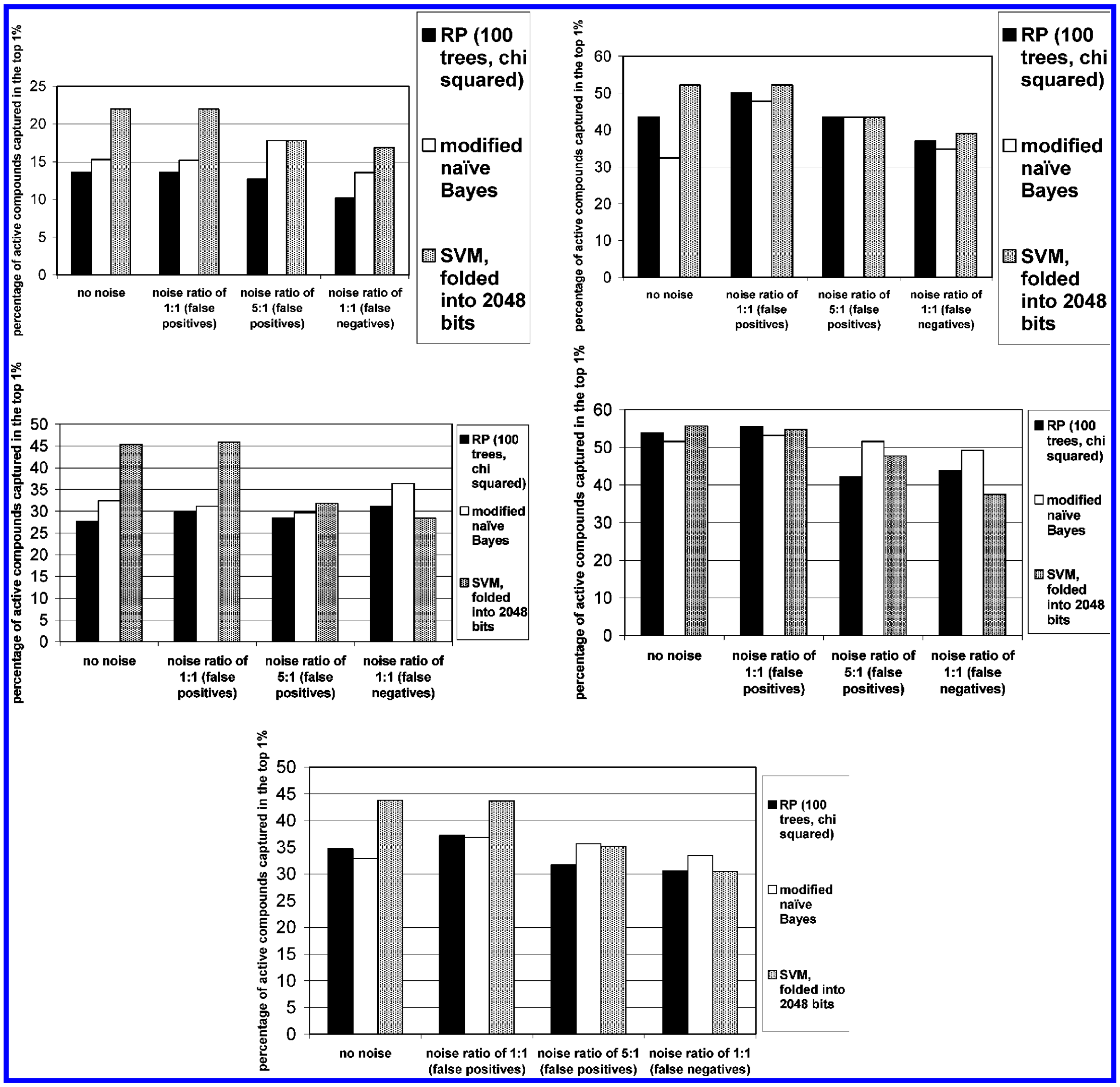

Figure 2. (a) % active compounds captured in the first 1% at increasing levels of noise for data set 1, (b) data set 2, (c) data set 3, and (d) data set 4. (e) The average % active compounds captured in the first 1% at increasing levels of noise for data sets 1−4.

on 100 trees and χ squared as the splitting criteria, Laplacian-modified naive Bayesian, and a SVM. For the SVM model, the 2048-bit folded fingerprint was utilized.

Since screening collections are designed for both chemical and biological diversity, the hit rate in a typical HTS campaign is low (ca. 0.1−1%). The active compounds—either true positives or false negatives—will, therefore, have a low occurrence and be a minority class. By employing diversity selection such as that suggested in this study on the ∼99% inactive compounds, a significant number of false negatives can be removed from the training set. Nevertheless, we studied the impact of false negatives, by picking 50% of the active compounds at random from the training set and changing their attribute to “inactive”. For each of the four data sets, three classifiers were generated: Laplacian-modified naive Bayesian, SVM, and recursive partitioning based on 100 trees and χ squared as the splitting criteria.

In terms of false positives, increasing the level of noise from no noise to a noise ratio of 1:1 had little impact on the accuracy of the classifiers in data sets 1, 3, and 4 as shown in Figure 2a−d. Strangely, in data set 2 (Figure 2b), increasing the noise to a ratio of 1:1 improved the percentage of compounds captured in the top 1% by the Laplacian-modified naive Bayesian model from 32.4% to 47.8%. When the noise was increased to a ratio of five misclassified compounds to one true active, the Laplacian-modified naive Bayesian and RP showed tolerance to the noise in data sets 1−3. In data set 4, the enrichment for the RP model declined from 53.9% to 42.2%. On average, SVM outperformed the other methods when the training sets included no noise or a noise ratio of 1:1 (Figure 2e). The performance of SVM was nearly identical to that of the Laplacian-modified naive Bayesian model with a noise ratio of 5:1.

In terms of false negative, employing noise in the ratio of 1:1 (equal number of true positives and false negatives in the training set) had little impact on the accuracy of the Laplacian-modified naive Bayesian model. Strangely, in data set 3, the percentage of compounds captured in the top 1% increased from 32.4% to 36.4% when false negative noise was employed. SVM seemed to be the most sensitive method to false negatives, where on average the percentage of compounds captured in the top 1% declined from 43.8% to 30.5% in the presence of false negatives (Figure 2e).

Data Set Scaffold Diversity. To get a better insight into the differences between the data sets and, particularly, to understand why the models were less successful on data set

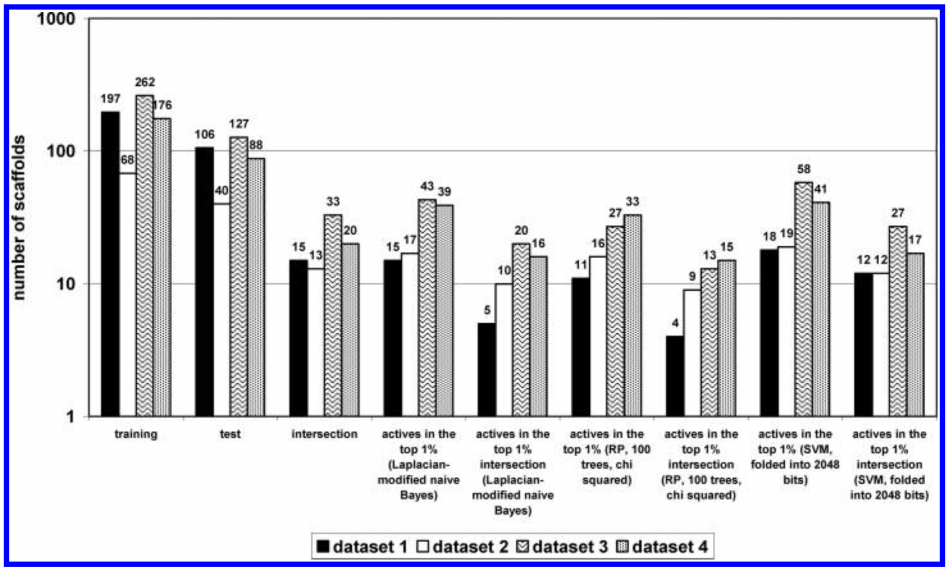

Figure 3. Number of Murcko scaffolds for the active compounds in both training ("training" column) and test sets ("test" column) for data sets 1−4. Since the data is distributed over 2 orders of magnitude, the *y* axis is presented in logarithmic units. The number of scaffolds appears on the top of each bar. The "intersection" column depicts the scaffolds in the test set that also appear in the training set. The number of scaffolds for the fraction of actives in the top 1% and the number of scaffolds for the fraction of actives in the top 1% that also appear in the training set are depicted for each of the three data mining methods.

1, the frequency of Murcko scaffolds[33] for the active compounds in both training and test sets were calculated as shown in Figure 3. Each different scaffold contributes one unit to the *y* axis; that is, a scaffold that appears *N* times is still counted as one. The "training" and "test" bars refer to the training and test sets, respectively, for each of the four data sets. The "intersection" bar depicts the scaffolds in the test set that appeared in the training set. For example, 15 out of the 106 (or 15/106 = 14%) scaffolds in the test set for data set 1 also appear in the training set. On the basis of Figure 3, the percentages of scaffolds in the test set that also appear in the training set for data sets 1, 2, 3, and 4 are 14%, 32%, 26%, and 22%, respectively. To identify which data mining method is better at finding "novel" scaffolds, we studied the number of scaffolds captured in the top 1% by the Laplacian-modified naive Bayesian, RP, and SVM classifiers in the four data sets. SVM outperformed the other two methods in capturing scaffolds of active compounds in the top 1%, where it captured 18, 19, 58, and 41 scaffolds for data sets 1, 2, 3, and 4, respectively ("actives in the top 1%", Figure 3). All the data mining methods could identify active scaffolds that did not appear in the training set. For example, in data set 1, the SVM model captured 18 scaffolds ["actives in the top 1% (SVM, folded into 2048 bits)", Figure 3]; six of them were "novel", that is, did not appear in the training set.

DISCUSSION

We have shown that the three data mining methods at hand, Laplacian-modified naive Bayesian, RP, and SVM, were extremely tolerant to stochastic noise. Indeed, when using a ratio of 1:1 of active and misclassified compounds in the training set, we did not observe any significant decline in the model's accuracy. SVM seemed to be the most sensitive method to noise in terms of percent change in enrichment compared to a model built from non-noisy data, particularly when the training set was contaminated with false negatives. It should be noted that, in three out of the four data sets, SVM still yielded the best enrichment in a noise ratio of 1:1 (false positives) compared with the other two methods and outperformed RP in a noise ratio of 5:1. In terms of false negatives, SVM still yielded the best enrichment in a noise ratio of 1:1 (equal number of false negatives and true positives in the training set). On average, the Laplacian-modified naive Bayesian classifier had the highest tolerance to false negatives and a high number of false positives (noise ratio of 1:5). Wiesenfeld and Moss[34] described a counterintuitive phenomenon defined as "stochastic resonance" where the addition of an appropriate amount of noise can actually improve the signal-to-noise ratio in a nonlinear system. This phenomenon has been observed in a variety of physical systems and even in biological sensory neurons, suggesting that it may be useful to animals.[35] It would be of value to understand if this phenomenon occurred in data set 2 (Figure 2b) where adding false positives in the ratio of 1:1 improved the percentage of compounds captured in the top 1% by the Laplacian-modified naive Bayesian model or data set 3 (Figure 2c) when the training set was contaminated with false negatives.

By comparing the Laplacian-modified naive Bayesian models with or without folding the fingerprint into 2048 bits, it is clear that there is a loss of information due to the folding. The Laplacian-modified naive Bayesian models based on the nonfolded fingerprint captured more active compounds and chemotypes (Murcko scaffolds) in the first 1%, and the area under the ROC curve values were higher in all four data sets. In terms of folding the fingerprint into 2048 bits, the added value of SVM outweighs the loss of information, particularly when looking into the number of active compounds captured in the first 1%. The nonfolded fingerprint method for SVM can reduce the loss of information over the folded one.

There is no strong correlation between the enrichment and ROC values. For example, SVM outperformed the Laplacian-modified naive Bayesian and RP classifiers in capturing active compounds in the top 1% in all four data sets. However, in terms of the area under the ROC curve, SVM was not the most robust data mining method for three out of the four data sets. This may be due to folding the fingerprints, which might have introduced additional noise and a loss of information. It could be argued that to capture active compounds in the first 1% is more important than the area under the ROC curve because subsequent validation or retesting of active hits will be limited to a few hundred.

On the basis of the Murcko scaffold analysis, one would, therefore, presume that data set 1 will be the most difficult to enrich since 85.8% of the scaffolds in the test set do not appear in the training set and that data set 2 will be the easiest to predict because of the presence of 32.5% of the test scaffolds in the training set. One would, therefore, expect the success of the models for each of the data sets to be in the following order: data set 2 > data set 3 > data set 4 > data set 1. In reality, the order for the area below the ROC curves was data set 2 > data set 3 ≈ data set 4 > data set 1, and for the enrichments, data set 4 > data set 2 > data set 3 > data set 1. The Murcko scaffold analysis, therefore, correlated well with the ROC values. The correlation with the percent of active compounds captured in the first 1% was less strong. Unlike data sets 2−4, which were generated from biochemical assays, data set 1 was created from a cell-based assay and it could be claimed that cell-based assays are considered noisier than biochemical assays, and therefore, data set 1 was the most difficult to enrich.

CONCLUSIONS

Mining HTS data remains inherently empirical but can have a number of important applications in the analysis of HTS results. Apart from the selection of compounds that are structurally related and that demonstrate some structure−activity relationships, or deprioritization of compounds which might have appeared as hits due to random noise, the development of models to describe HTS data can also be used in subsequent screening efforts of focus sets or to identify compound classes that had not previously been tested.

We found that a well-tuned SVM model is an attractive avenue for enriching data sets because of its high accuracy when the training set is relatively "clean" from noise. All three methods performed well on noisy data sets. The Laplacian-modified naive Bayesian model relies on the simplistic and naïve assumption that the features are independent. It, however, works surprisingly well, particularly in high levels of noise. We also find it attractive since the computing time scales linearly with the number of points and it does not require a significant amount of tuning by the user.

ACKNOWLEDGMENT

The authors express their appreciation to Dr. Christian Parker and Dr. Greg Paris for their helpful feedback and discussions.

REFERENCES AND NOTES

(1) Rusinko, A., III; 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.

(2) Karnachi, P. S.; Brown, F. K. Practical approaches to efficient screening: information-rich screening protocol. *J. Biomol. Screening* **2004**, *9*, 678−686.

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

(4) Glick, M.; Klon, A. E.; Acklin, P.; Davies, J. W. Enrichment of extremely noisy high-throughput screening data using a naive Bayes classifier. *J. Biomol. Screening* **2004**, *9*, 32−36.

(5) Godden, J. W.; Bajorath, J. An information-theoretic approach to descriptor selection for database profiling and QSAR modeling. *QSAR Comb. Sci.* **2003**, *22*, 487−497.

(6) Wu, X.; Glickman, J. F.; Bowen, B. R.; Sills, M. A. Comparison of assay technologies for a nuclear receptor assay screen reveals differences in the sets of identified functional antagonists. *J. Biomol. Screening* **2003**, *8*, 381−392.

(7) Kelley, B. P.; Lunn, M. R.; Root, D. E.; Flaherty, S. P.; Martino, A. M.; Stockwell, B. R. A flexible data analysis tool for chemical genetic screens. *Chem. Biol.* **2004**, *11*, 1495−1503.

(8) Diller, D. J.; Hobbs, D. W. Deriving knowledge through data mining high-throughput screening data. *J. Med. Chem.* **2004**, *47*, 6373−6383.

(9) Hastie, T.; Tibshirani, R.; Friedman, J. In *The Elements of Statistical Learning − Data Mining, Inference and Prediction*; Springer: New York, 2005.

(10) Xia, X.; Maliski, E. G.; Gallant, P.; Rogers, D. Classification of kinase inhibitors using a Bayesian model. *J. Med. Chem.* **2004**, *47*, 4463−4470.

(11) Vapnik, V. In *Statistical Learning Theory*; Wiley: New York, 1998.

(12) Klon, A. E.; Glick, M.; Davies, J. W. Application of machine learning to improve the results of high-throughput docking against the HIV-1 protease. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2216−2224.

(13) Klon, A. E.; Glick, M.; Davies, J. W. Combination of a naive Bayes classifier with consensus scoring improves enrichment of high-throughput docking results. *J. Med. Chem.* **2004**, *47*, 4356−4359.

(14) Blower, P.; Fligner, M.; Verducci, J.; Bjoraker, J. On combining recursive partitioning and simulated annealing to detect groups of biologically active compounds. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 393−404.

(15) Tong, W.; Hong, H.; Fang, H.; Xie, Q.; Perkins, R. Decision forest: combining the predictions of multiple independent decision tree models. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 525−531.

(16) Shi, L. M.; Fang, H.; Tong, W.; Wu, J.; Perkins, R.; Blair, R. M.; Branham, W. S.; Dial, S. L.; Moland, C. L.; Sheehan, D. M. QSAR models using a large diverse set of estrogens. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 186−195.

(17) Warmuth, M. K.; Liao, J.; Ratsch, G.; Mathieson, M.; Putta, S.; Lemmen, C. Active learning with support vector machines in the drug discovery process. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 667−673.

(18) Witten, I. H.; Frank, E. In *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*; Morgan Kaufmann Publishers: San Francisco, CA, 1999.

(19) SciTegic homepage. http://www.scitegic.com.

(20) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. In *Classification and Regression Trees (CART)*; Wadsworth: Pacific Grove, CA, 1984.

(21) Quinlan, J. R. Improved use of continuous attributes in C4.5. *J. Artif. Intell. Res.* **1996**, *4*, 77−90.

(22) Loh, W. Y.; Vanichsetakul, N. Tree-Structured Classification Via Generalized Discriminant-Analysis. *J. Am. Stat. Assoc.* **1988**, *83*, 715−725.

(23) Brodley, C. E.; Utgoff, P. E. Multivariate Decision Trees. *Machine Learning* **1995**, *19*, 45−77.

(24) NovoDynamics, Inc. homepage. http://www.novodynamics.com.

(25) Equbits, LLC. homepage. http://www.equbits.com.

(26) Dixon, S. L.; Merz, K. M., Jr. One-dimensional molecular representations and similarity calculations: methodology and validation. *J. Med. Chem.* **2001**, *44*, 3795−3809.

(27) Brown, R. D.; Martin, Y. C. Use of structure activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.

(28) Matter, H.; Potter, T. Comparing 3D pharmacophore triplets and 2D fingerprints for selecting diverse compound subsets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1211−1225.

(29) Schuffenhauer, A.; Gillet, V. J.; Willett, P. Similarity searching in files of three-dimensional chemical structures: Analysis of the BIOSTER database using two-dimensional fingerprints and molecular field descriptors. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 295−307.

(30) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Org. Biomol. Chem.* **2004**, *2*, 3256−3266.

(31) Morgan, H. L. Generation of a unique machine description for chemical structures−a technique developed at chemical abstracts service. *J. Chem. Soc.* **1965**, *5*, 107−113.

(32) Schuffenhauer, A.; Popov, M.; Schopfer, U.; Acklin, P.; Stanek, J.; Jacoby, E. Molecular diversity management strategies for building and enhancement of diverse and focused lead discovery compound screening collections. *Comb. Chem. High Throughput Screening* **2004**, *7*, 771−781.

(33) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.

(34) Wiesenfeld, K.; Moss, F. Stochastic resonance and the benefits of noise: from ice ages to crayfish and SQUIDs. *Nature* **1995**, *373*, 33−36.

(35) Russell, D. F.; Wilkens, L. A.; Moss, F. Use of behavioural stochastic resonance by paddle fish for feeding. *Nature* **1999**, *402*, 291−294.

CI050374H