

Creating a Smart Virtual Screening Protocol, Part II: Recursive Partitioning for Sequential Screening

Shikha Varma*, Luke Fisher, Teresa Lyons, Deqi Chen

For further information or inquires, please contact Shikha Varma, shikha@accelrys.com

Abstract

This study presents a sequential screening workflow that was developed based on data from a high throughput screening (HTS) of a diverse, 50,000 compound library containing only 32 active inhibitors of *Escherichia coli* dihydrofolate reductase (DHFR). The workflow employs raw data analysis and recursive partitioning (RP), followed by docking and scoring, to significantly reduce the data set and improve enrichment over docking and scoring alone. The methods used are rapid and work easily with large data files and interpretable descriptors to remove many inactive compounds from a dataset. When used in this instance, the workflow eliminated many false positives prior to virtual HTS/scoring. This substantially decreased the amount of data analysis required and led to almost half of the active compounds being represented in the top 1% of the ranked compounds.

Introduction

The "what" and "why" of sequential screening

Pharmaceutical companies can use experimental data from HTS assays to develop *in silico* models that are used to create focused libraries for further HTS. The process of pre-filtering or reducing false positives by using such *in silico* models is normally termed 'sequential screening.' The purpose of sequential screening is to find better compounds and remove failure compounds at an early stage. This is done by using a training set of compounds from an initial screen and building a model that differentiates between high-activity and low-activity compounds. As shown in Figure 1, *in silico* screening of a compound collection (commercially available, corporate, or virtual libraries) against the resulting model helps prioritize the next iteration of screening and model building. This approach can accelerate the rate of drug discovery and save costs by providing a greater probability of finding a lead compound.¹⁻³

Industry Sector

Pharmaceutical

Organization

Accelrys

Key Products

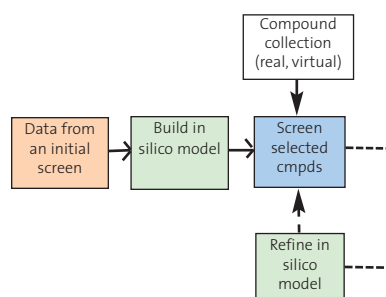
DIVA

Cerius²-CSAR

Cerius²-Descriptor+

Cerius²-LigandFit

Cerius²-LibDock



▲ Figure 1: Sequential screening paradigm.

Background

Which statistical algorithm to use?

There are many statistical approaches to build theoretical models that attempt to correlate structures of compounds to their observed activities. The choice of which algorithm to use depends upon the quantity of the data, quality of the response, and the type of the response being measured (numerical vs. categorical data). A vast amount of literature is available that overviews statistical algorithms for correlating *quantitative* structure-activity relationships (QSAR) of a dataset. However, with quantitative, *qualitative*, and *categorical* activities, only classification algorithms that attempt to form a model by dividing a dataset into mutually exclusive groups can be used. Such classification-activity relationship models define how similar or different these compounds are to each other with respect to their variables. There are a number of algorithms that can be used for classification of data, including Neural Nets (NN), linear discriminant analysis (LDA), soft independent modeling of class analogy (SIMCA), K nearest neighbor (KNN), classification regression tree (CART), etc. All of these classification methods, including several others, were recently compared for their performance on a toxicity dataset. The authors of this study concluded that the CART algorithm performed best overall as a classification method on the dataset at hand.⁴ There are many variations of the CART algorithm. This study focuses on its implementation in the Cerius² molecular modeling package.

Recursive partitioning

The CART methodology involves RP of a dataset matrix to make a hierarchical decision tree (rows = compounds and columns = descriptors, X and activities, Y).^{5,6} The decision tree is constructed by one question per node, creating a binary (yes/no) split and resulting in two statistically distinct subsets or nodes. The splitting criteria are determined by a statistical analysis of each variable and the assigned categorical activity of the compounds. The splitting process continues until no more significant nodes are obtained or when a minimum number of samples per node are reached (a specified parameter) and a class prediction is made on the terminal node. Another parameter in the algorithm prunes the tree to the appropriate tree depth. In the current implementation of RP in Cerius², this algorithm is enhanced and termed 'Lookahead RP,' which attempts to provide all possible trees up to the level of complexity specified, finally leading to a more predictive tree after pruning.

Besides being able to correlate categorical as well as numeric activities, RP offers many other advantages for classifying datasets. For example, RP accounts for non-linear relationships and all variables are considered at every stage of model building. During the variable selection process, RP takes into account the prior probabilities and penalties for misclassification. RP can also use any type of compound variables, including 2D and 3D fingerprints. Moreover, there are many splitting options available to classify highly skewed or unbalanced datasets, such as data from *in vitro*/*in vivo* HTS that consist of very few actives and mostly inactives compounds.

Recent advances in one of the RP algorithms in Cerius² RP accommodate multiple dependent properties; this algorithm is termed Partially Unified Multiple Property Recursive Partitioning or PUMP-RP.^{7,8} With it, a separate activity class can be predicted for each of several targets with a single tree model.

From a practical standpoint, there are several advantages to RP. For instance, unlike typical QSAR algorithms, RP can handle datasets containing hundreds to thousands of compounds. The RP models are fast to build and the output provides a graphical tree, as well as a statistics table that includes correctly classified percentages, the number of false positives, and compound member information for each node. It is easy to interpret which descriptors are more important than others from the final RP model. Applications of Cerius² RP and PUMP-RP in a pharmaceutical environment range from HTS data analysis^{2,3} to ADME model generation⁹ to understanding drug aggregate formation in HTS.¹⁰

HTS Dataset

To illustrate the concept of sequential screening, data from an *in vitro* HTS of a diverse library of 49,995 small molecules against *E. coli* DHFR was used.¹¹ DHFR is a relatively small protein whose activity is critical for DNA synthesis. It is the target for at least four drugs on the market whose benefits range from antibiotic to anticancer. The protein uses a cofactor termed NADPH to accomplish the reduction of dihydrofolate to tetrahydrofolate. Attempts with receptor-based virtual HTS have been a challenge because the binding pocket is large; therefore, a number of ligand-based approaches have been made to identify more novel and potent DHFR inhibitors.

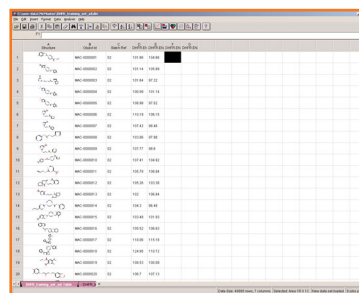
In the present dataset, activity against DHFR was collected in replicates for each compound. Actives were defined as those compounds that affected DHFR activity to 75% residual activity or less in both replicates. Thirty-two compounds were classified as active after the primary screen.

Methodology

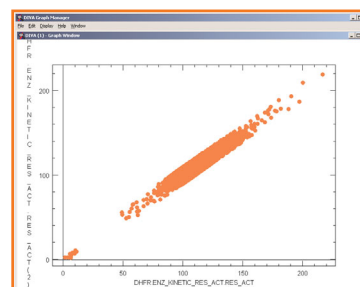
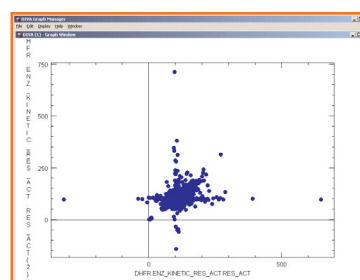
Examining the raw data

The first step was to examine the raw data from the primary screen. It is a well known fact that HTS screening data often carries outliers. Prior to using such activity data to build an *in silico* model, it is absolutely necessary to eliminate data points that are not consistent across replicates. As shown in Figures 2 and 3, this was easily done by reading the raw structure data file (.sdf) on a desktop application, DIVA, and plotting the two sets of activities against each other.

As shown in Figure 3, the HTS screen data consists of a number of outliers, meaning that the duplicate data points were not consistent in the two assays. This often results from different assay conditions, temperature differences, tilted assay plates, etc. Compounds were removed from the dataset if the assay values were more than 10% different. This is a subjective choice and entirely depends upon the biological assay being considered. In this case, after removal of the outliers, the final dataset consisted of 43,195 compounds. This



▲ Figure 2: Raw HTS structure data files can be easily read into DIVA for data visualization and analysis.



▲ Figure 3: (Top) A plot of the HTS screen data; the two duplicate assay points are plotted against one another. (Bottom) The same duplicate assay data points plotted after removal of outliers.

set was subsequently verified for the presence of 32 known actives. Thus, 6,768 inactive compounds were removed from the original dataset.

Building the RP model

Prior to building an RP model in Cerius², the following steps need to be carried out:

- The data need to be converted to a binary file format (BDF) because working through a study table is not amenable due to the size of the data.
- Numerical percentage activities have to be 'translated' into categories (e.g., high, low, medium, etc). In this case, we use LOW to annotate active compounds and HIGH to annotate inactive compounds.
- A set of descriptors need to be selected and calculated for all the compounds.
- The RP parameters have to be optimized.

In Cerius², the BDF capability allows one to use large amounts of data in a single file, calculate properties through the FastDescriptors facility, and perform statistical analysis. A total of 84 descriptors were calculated, including E-state keys, electronic and information content, spatial, structural, thermodynamic, and topological. Some of these descriptors were ignored in the final model building due to zero variance. Finally, categorical activities were associated with each compound as follows:

- If the activity was measured <75%, the category was assigned LOW (i.e. Active).
- If the activity was measured >75%, the categorical activity was assigned as HIGH (i.e. Inactive).

Parameters

The following parameters were optimized to get the final RP model:

- Type of weighting by: Classes.
 - Each activity class is treated equally rather than each compound.
- Score splits using: Gini Impurity.
 - This is type of scoring function that determines how sub-groups should be divided statistically.
- Pruning factor: 343.
 - Pruning factor determines the tree depth at each node. This number was achieved by first running RP with moderate pruning and looking at the alpha values in the statistics. Then a tree is chosen based upon one's objective.
- Minimum number of samples per node: 5.
- Lookahead node must contain: 5 samples.
- Limit knots per variable: 50.
 - Number of ways a descriptor is evaluated for its statistical significance.
- Maximum Lookahead node depth: 5.
 - A number other than 0 turns on the Lookahead RP method; it specifies the maximum tree depth at which alternative splits for a node are followed past immediate children.

Table 2 highlights the selected RP model for this dataset with the following parameters:

- Value: Actual class memberships defined as the dependent (Y) column.
- #: Number of samples in each class.
- %: Percentage of compounds in each class.
- Class %ObsCorrect: Intra-class prediction.
 - Only the molecules in the corresponding class are being predicted. It provides information on false negatives as well as false positives, depending on which class you want to examine.
- Overall %PredCorrect: Overall prediction.
 - In this test, all the molecules in the set are being predicted. This provides you with some information on the accuracy of the prediction when you predict the whole set with the model.
- Enrichment: The enrichment factor for a specific bin is the percentage of compounds correctly predicted to belong to that bin (Overall %PredCorrect) divided by the original percentage of compounds belonging to that bin (%).

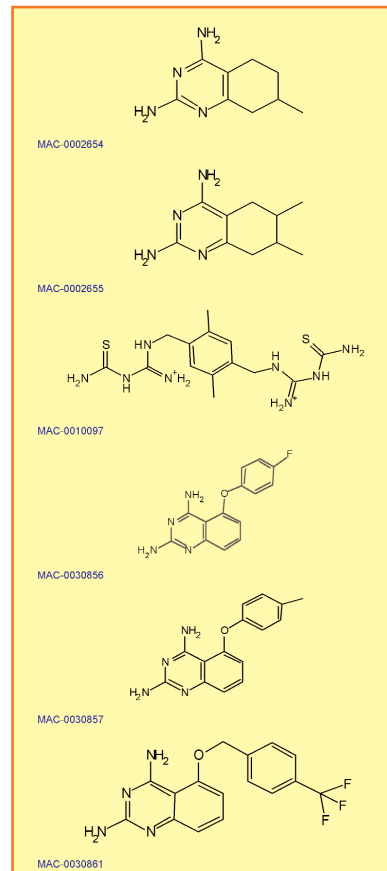
	Value	#	%	Class %ObsCorr	Overall %PredCorr	Enrichment
1.00	HIGH	43195	99.93	91.19	99.99	1.00
2.00	LOW	32	0.07	93.75	0.78	10.57

▲ Table 2: Statistics table produced for the final RP model.

Conclusions

An enrichment of over 10-fold was observed with the 14-leaf model; that is, if one were to run RP on the 50K compound dataset by screening only 3,804 compounds, 30 out of 32 actives could be recovered. A 14-leaf model was selected in this study because, upon examining the data table of alpha values, it was found that the tree would need to have 25 leaves in order to recover all the actives (data not shown). Therefore, nine more leaves were required for correctly classifying two more actives. The decision of which model to select is strictly a matter of choice. For example, one might argue that a 25-leaf RP model could lead to an over trained model in order to retrieve two more actives. Once again, it depends on the objective of constructing such an RP model. If the objective is to reduce false positives in an experimental HTS, then such a 14-leaf *in silico* model can help reduce the number of false positives considerably. Moreover, this model has an extremely high enrichment for classifying compounds with primary amine groups, as shown in Figure 7. Therefore, if the objective is to look for compounds with diverse chemistry space of DHFR inhibitors, then the 25-leaf model can be used because it is able to retrieve all the actives.

Using this RP model, the predicted set of 3,804 compounds was then virtually screened with Cerius²·LigandFit (Site Features Docking) and the compounds were prioritized on the basis of various scoring functions. The results



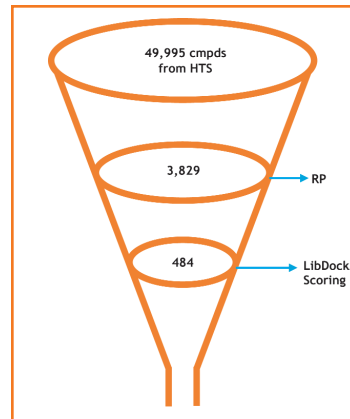
▲ Figure 7: Some representative DHFR inhibitors containing primary amine groups.

from this screening were then compared to the virtual HTS docking and scoring of the raw dataset of all 50K compounds. As shown in Figure 8, by eliminating false positives with an RP screen, almost a half of the active compounds were represented in the top 1% of the ranked compounds. The details of docking and scoring are presented more elaborately in a separate study.¹²

In summary, RP provides a smart way to pre-screen primary HTS data by eliminating false positives and it improves enrichment by as much as 10-fold. Furthermore, it is also very important to examine the raw data prior to building *in silico* models and performing virtual HTS. The descriptors provided for building RP models also provide meaningful insight into structure-activity relationship in the dataset.

References

- 1) Engles M.F.M., Thielemans T., Verbinnen D., et al, "CerBeruS: A System Supporting the Sequential Screening Process," *J. Chem. Inf. Comput. Sci.*, **2000**, *40*, 241-245.
- 2) van Rhee A.M, Stocker J., Printzenhoff D., et al., "Retrospective Analysis of an Experimental High-Throughput Screening Data Set by Recursive Partitioning," *J. Comb. Chem.*, **2000**, *3*, 267-277.
- 3) van Rhee A.M.. "Use of Recursion Forests in the Sequential Screening Process: Consensus Selection by Multiple Recursion Trees," *J. Chem. Inf. Comput. Sci.*, **2003**, *43*, 941-948.
- 4) Mazzatorta P, Benfenati E, Lorenzini P, and Vighi M., "QSAR in Ecotoxicity: An Overview of Modern Classification Techniques," *J. Chem. Inf. Comput. Sci.*, **2004**, *44*, 105-112.
- 5) Breiman L, Friedman J.H., Olshen R.A. and C.J. Stone, in *Classification and Regression Trees*, Wadsworth, **1984**.
- 6) Hawkins D.M., Young S.S. and Rusinko III A., "Analysis of Large Structure-Activity Data Set Using Recursive Partitioning," *Quant. Struct.-Act. Relat.*, **1997**, *16*, 296-302.
- 7) Stockfisch T.P., "Partially Unified Multiple Property Recursive Partitioning (PUMP-RP): A New Method for Predicting and Understanding Drug Selectivity," *J. Chem. Inf. Comput. Sci.*, **2003**, *43*, 1608-1613.
- 8) Rao S.N. and Stockfisch T.P., "Partially Unified Multiple Property Recursive Partitioning (PUMP-RP) Analyses of Cyclooxygenase (COX) Inhibitors," *J. Chem. Inf. Comput. Sci.*, **2003**, *43*, 1614-1622.



▲ Figure 8: Number of compounds reduced for data analysis through sequential screening.

9) O'Brien, S. E. and de Groot M., Pfizer, Sandwich, UK. Oral presentation, 227th ACS National Meeting, Anaheim, CA, April, **2004**.

10) Seidler J., McGovern S.L., Doman T.N., and Shoichet B.K., "Identification and Prediction of Promiscuous Aggregating Inhibitors among Known Drugs," *J. Med. Chem.*, **2003**, *46*, 4477-4486.

11) Zolli-Juran M., Cechetto J.D., Hartlen R., et al., "High Throughput Screening Identifies Novel Inhibitors of Escherichia coli Dihydrofolate Reductase that are Competitive with Dihydrofolate," *Bioorg. Med. Chem. Lett.*, **2003**, *13*, 2493-2496.

12) <http://www.accelrys.com/cases>