

Shikha Varma-O'Brien^a, David Rogers^b

^aAccelrys, Inc., 10188 Telesis Court, Suite 100, San Diego, CA.

^bScitegic, 10188 Telesis Court, Suite 100, San Diego, CA.

Abstract This work describes the application of Bayesian modeling, as implemented in PipelinePilot™, to 17,550 compounds and their corresponding cyclin-dependent kinase-2 (CDK2) activities¹. The molecules in the dataset have been classified into one of 22 scaffolds, some of which are generally more associated with CDK2 inhibition than others. Bayesian modeling provides an ideal way to rapidly analyze this data with a view to library development and compound prioritization. This model distinguishes good CDK2 inhibitors (actives) from the bad CDK2 inhibitors (inactives) by using Scitegic's proprietary chemical fingerprints, FCFP_6. These fingerprints enable us to recover chemical scaffolds and sub-structures that are intrinsically associated with CDK2 activity. Receiver Operating Characteristic (ROC) plot with the area under the curve (AUC) of 0.83, reveals the significant enrichment obtained using this virtual screening methodology: 17% of active compounds are identified by screening just 1% of the database.
¹Bradley, E. K. et al., *J. Med. Chem.* 2003, 46, 4360-4364.

Introduction

Cyclin-dependent kinases are cellular kinases which play a crucial role in phases of cell cycle. CDK2 inhibitors are being studied by many groups for their potential as anticancer therapeutics. Structure and ligand based methods are the two general categories of computational methods employed to assist in the process of rational discovery of active compounds. The present study is a ligand-based retrospective analysis of the classification of activity of ligands using a Bayesian statistical approach. This work demonstrates how CDK2 inhibitor-like libraries can be generated for smart screening using a model developed below. This model also allows identification of structural features that are associated with CDK2 activity.

CDK2 HTS Data

Data from previously published work are used in this study. The library contains a combination of sources of compounds: diverse compounds from general screening libraries (13,359 with 207 actives), commercial compounds selected by chemists (951 with 161 actives), and the rest synthesized /screened iteratively using 22 scaffold types (3,240 total with 14 actives) [1]. We combined the sources and divided the data equally and randomly for training and test sets.

- Total dataset of 17,550 CDK2 inhibitors.
- Compounds with an IC₅₀ lower than 25 nM or inhibition larger than 50% at 10 nM were considered active

Representative Compounds in Each Scaffold Class:

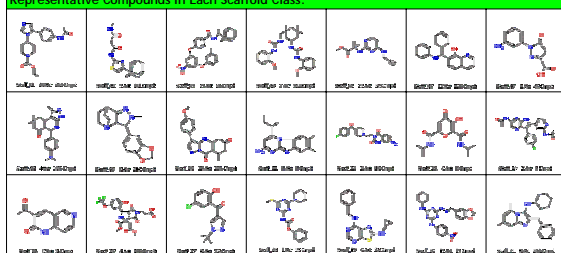


Figure 1. Example compounds from the CDK2 HTS data showing the number actives from the total compounds in each scaffold class. The smallest compound (fewest atoms) from each scaffold class is featured.

Method

We describe here the application of a rapid and powerful method of developing predictive models using Laplacian-modified Bayesian statistics available in the Scitegic product Pipeline Pilot™ with the proprietary functional class fingerprints (FCFP_6) as shown in Figure 2. Bayesian learning is ideal for vHTS applications because:

- Efficient: It is fast and it scales linearly with large data sets
- Robust: Works for a few as well as many 'good' (or eg. Active) examples
- Unsupervised: No tuning parameters needed
- Multimodal: Can model broad classes of compounds and multiple modes of action represented in a single model

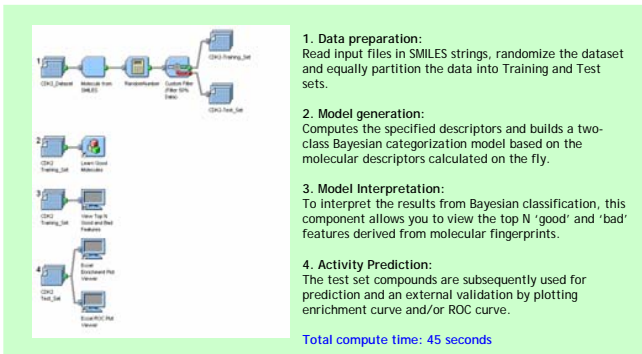


Figure 2. Pipeline Pilot™ Protocols

The descriptors used in this study are FCFP_6, ALogP, molecular weight, number of hydrogen bond donors, number of hydrogen bond acceptors, and number of rotatable bonds.

FCFP_6 are 2D fingerprints where the atom types are abstracted to the role that the atom plays in the molecule. The generation of an FCFP fingerprint for a molecule involves with the assignment of an atom code for each heavy (non-hydrogen) atom in the molecule. The atom code is based on an estimate of the functional role of the atom (HBA, HBD, positively ionized or positively ionizable, negatively ionized or negatively ionizable, aromatic, and halogen) and its neighbors.

A model is constructed through a learning process with Bayesian statistics, comparing the frequency of occurrence of fingerprint features contained in the actives molecules against the overall frequency in all molecules in the dataset. A count of each feature is kept over all the samples as well as the samples that pass the test for 'good' - a baseline probability. A normalized probability is then calculated for each feature as log(Laplacian corrected probability) and a relative score is determined by summing the normalized probabilities over all features. This estimator is used to adjust the uncorrected probability estimates of a feature to account for the different sampling frequencies of different features [2].

Figure 3 shows a truncated output statistics table from the Bayesian categorization. Each property that is used to build the model is listed along with its statistics. The total number of features from this property from all molecules is indicated (Total #), along with the total number of features from all molecules in the "good" subset. Each bin indicates the number of occurrences of a given feature in the compound. The table lists the Positive bin (where "positive" means the bin uses a feature that increases the likelihood that a sample is "good") as well as the Negative bin (where those features are found less frequently in the good compounds).

Once a model is built, every molecule is given a prediction score based on the contributions from each constitutive feature. This enables the compounds to be ranked in order of their probability of having CDK2 activity. In addition, each compound can be classified as "good" (active) or "bad" (inactive) depending on whether its score is greater than or less than a predetermined classification cut-off value. This value is the score that best divides the training set of compounds into their activity classes. In this model, a value of 0 provided the desired separation of molecules.

Figure 3. (see text)

Equation "CDK"									
Features from: ("FCFP_6" LongFingerprintType)									
Features from: ("ALogP" DoubleType)									
Features from: ("Molecular_Weight" DoubleType)									
Features from: ("Num_H_Donor" LongType)									
Features from: ("Num_H_Acceptor" LongType)									
Features from: ("Num_AromaticRings" LongType)									
Feature Statistics:									
Property: "FCFP_6"									
Total # of features in all samples: 451615 in subset: 11251									
POSITIVE BINS					NEGATIVE BINS				
Bin ID	G1	G2	G3	G4	Bin ID	B1	B2	B3	B4
Bin Value	-1.93E+09	77818758	-1.73E+09	1.47E+08	Bin Value	3.3E+08	-5.51E+08	4.94E+08	3.94E+08
Feature Count	74	54	56	57	Feature Count	341	338	325	290
Subset Count	23	17	17	17	Subset Count	0	0	0	0
Normalized Probe	2.153001	2.037962	2.016639	2.006052	Normalized Probe	-2.250794	-2.242391	-2.207908	-2.107144

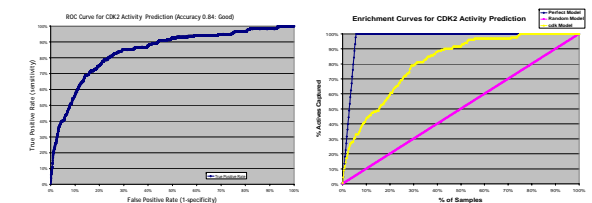
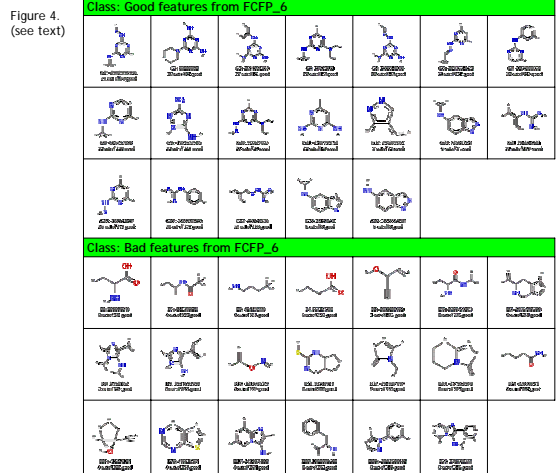


Figure 5. ROC curve (left) and the Enrichment plot (right) from the classification of test set CDK2 data.

Results: Predictions

The Bayesian model was able to correctly classify over 80% of the 8,706 test set compounds (129 out of 173 actives and 6,898 out of 8,533 inactives). This is a good prediction but a major strength of Bayesian modeling is its ability to rank compounds according to their probability of being active. This ranking is important when prioritizing compounds for screening or for further development. The Receiver Operating Characteristic (ROC) plot and the Enrichment curve featured in Figure 5 are graphical representations of the quality of this ranking. A ROC plot is a graph of sensitivity vs. specificity for different cutoff values, where:

- **sensitivity** is the ability of a model to avoid false-negatives. This is also called the 'true positive rate'.
- **specificity** is the ability of a model to predict true negatives. It is defined as the percentage of negative (non-good) samples (versus all true negatives) to the left of the cutoff. The actual plot axis is not specificity, but (1 - specificity), also called the 'false positive rate'. Therefore, a ROC plot can be thought of as a plot of the true positive rate vs. the false positive rate. The area under the curve represents accuracy, 1.0 being a perfect model [3].

The enrichment curve plots the number of active compounds recovered versus the proportion of the database screened. The pink diagonal line shows how many active compounds would be recovered when the database is screened randomly. This graph shows that by screening just 1% of the database, 17% of the actives are retrieved (17-fold enrichment).

Both of these graphs demonstrate the quality of the Bayesian model generated for the prediction of CDK2 inhibition, in particular the ability of the model to accurately rank compounds according to activity.

Results: Interpretation

Figure 4 shows the output from the "View Top N Good and Bad Features" component; it pulls out the top good and bad features derived from FCFP_6 (and other) descriptors from this model and indicates its frequency associated with a good compound. In this case the number of features required was set to 20 and displays only a single example of a particular feature. For example, the first cell indicates that FCFP_6 fingerprint representing that substructure, G1, was found 74 times and out of those it occurred in 23 good compounds.

A few conclusions can be derived from this table: the fingerprint features most closely associated with activity are shown as G1-G5. Their normalized probability scores range between +2.13 and +2.01. All 15 test set compounds belonging to scaffold_20 contain FCFP_6 features G1-G5 and are correctly predicted to be active.

As previously reported [1], scaffold_05 was designated as an "active" scaffold. Our model recovers 21 out of the 23 actives within the test set. This is because scaffold_05 contains the sub-structure shown in G11. This structure was seen 52 times in the training set, 13 times in active compounds giving it a normalized probability score of +1.81. However, G11 also occurs in the inactive molecules belonging to scaffold_05. As this scaffold itself is deemed to confer activity, 112 scaffold_05 compounds are incorrectly predicted to be active (false positives).

The model shows that FCFP_6 features B16 and B17 (normalized probability scores of -1.68 and -1.63 respectively) were never associated with good CDK2 activity. The test set compounds with Scaffold_19 and Scaffold_21 contained these functional groups and hence were correctly predicted to be inactive.

Conclusions

- Model was built with thousands of compounds with diverse structures and scaffolds
- Model generation extremely fast (in seconds) and easy in Pipeline Pilot™
- Model enables to correlate structural features with biological activities
- Correctly classified compounds
- Ranked compounds providing significant enrichment
- Scaffolds conferring CDK2 inhibitory activity were identified

References:

- [1] Bradley, E.K., Miller J.L., Salah, E. and Grootenhuys, P.D.J. *J. Med. Chem.*, 2003, 46, 4360-4364.
- [2] Xia, X., Maliski, E. G., Gallant, P. and Rogers, D. *J. Med. Chem.*, 2004, 47, 4463-4470.
- [3] Scitegic, Pipeline Pilot v.5.0 Help guide, 2005