



Bayesian Modeling

Bayesian modeling, as implemented in Pipeline Pilot, is a two class learner that builds a model to predict the likelihood that a given data sample is from a "good" subset of a larger set of baseline samples.

In application to HTS analysis, this means that a model will be learned of good hits from a baseline of inactives from the screen. Once learned, the model can be applied to a set of molecules whose activity is unknown and provides a score whose value gives a prediction of the likelihood that the molecule will be a hit. Bayesian statistics has a number of characteristics for the analysis of HTS data:

1. Efficient: models are extremely fast to calculate, even for hundreds of thousands of samples. The algorithm is single-pass and scales linearly with data set size.
2. Robust: works well for few samples of good (e.g. a very low hit rate typical of HTS) as well as for many samples of good.
3. Unsupervised: no tuning parameters are needed.
4. Multimodal: multiple modes of action in a screen may be represented in a single model.

In general the Bayesian modeling of activity data consists of the following steps:

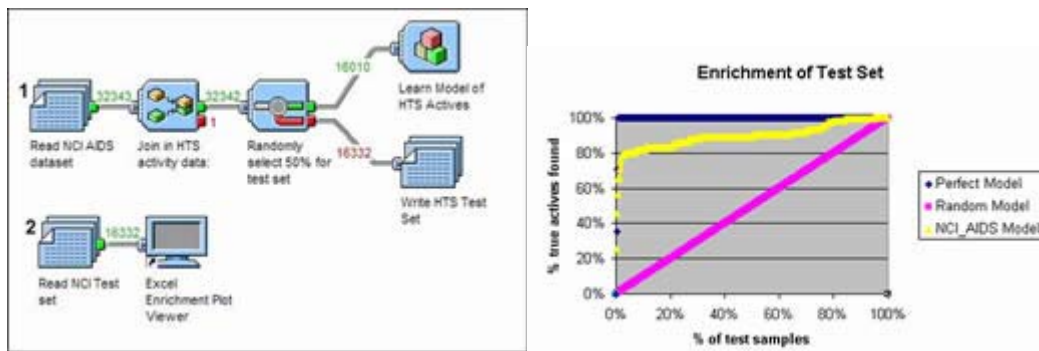
1. For the training data of molecules with known activity, calculate a set of descriptors, typically chemical fingerprints or physicochemical properties
2. Apply the Bayesian statistics to assign the probability, for each individual descriptor (fingerprint bit or property range), of a molecule's likelihood to be a member of the good class, given the presence of the descriptor. The table below shows the statistics for the most discriminating features in an example model. It shows the number of molecules in which the feature occurred, and the number of those molecules that were active. The Bayesian score is a measure of how different this is from the hit rate as a whole - which is the ratio that would be expected if the features is occurring at random across the actives and inactives. The score also takes account of the total number of occurrences of the feature, ensuring more weight is placed on features that are seen more often and little weight on those for which there are very few occurrences.
3. Validate the model. The component automatically performs an extremely fast leave-one-out cross validation as the model is built, providing a variety of statistics to allow assessment of the model's quality.
4. Apply the model to make predictions.

Good features from FCFP_6				
 01 - 42679621 15 out of 20 good Bayesian Score 2.460	 02 - 821107162 15 out of 19 good Bayesian Score 2.329	 03 - 438918276 11 out of 52 good Bayesian Score 2.315	 04 - 66452775 11 out of 16 good Bayesian Score 2.283	 05 - 41098826 15 out of 16 good Bayesian Score 2.283
Bad features from FCFP_6				
 06 - 164942021 0 out of 262 good Bayesian Score -1.908	 07 - 294124776 0 out of 525 good Bayesian Score -1.822	 08 - 108008826 0 out of 524 good Bayesian Score -1.819	 09 - 41240826 1 out of 699 good Bayesian Score -1.814	

Validation

The Bayesian learning has been extensively validated and a number of studies on its use have been published. Here we present a simple analysis of the NCI AIDS data set to show its use. The data set contains 32,000 compounds selected for HTS in a whole

cell HIV assay. 230 were confirmed as hits. A simple learning experiment is performed to separate the data into equal size training and test sets, learning on the former and then predicting the latter.



The enrichment curve shows the test set sorted by Bayesian score on the X-axis and the number of actives recovered as that list is traversed on the Y-axis. It shows that, if the test set had not already been screened, screening only 4% of the samples (600 compounds out of 16000) with the highest scores would have been sufficient to recover over 80% of the activity data of the whole set. In a real life screening situation this would represent a considerable cost saving in terms of time and materials.

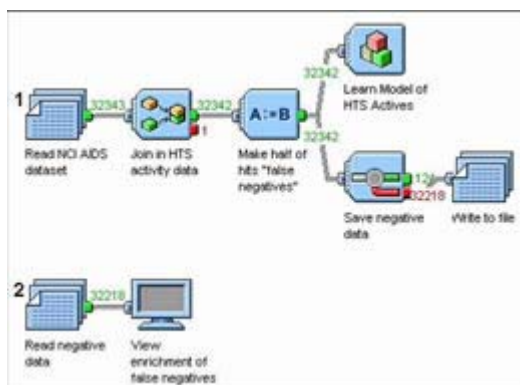
Applications

Bayesian models learned from HTS data have been applied in a number of use cases. The simplest case, such as that shown above, is to build a model from the results of a screening run, use the model to predict the activity of molecules proposed for synthesis or acquisition, and then only make or buy a small subset based on their predicted activity.

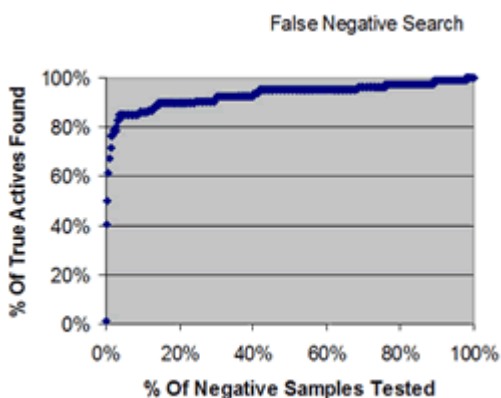
Learning has also been used in the search for false negatives - molecules that test as inactive but are in fact hits. False positives are typically discovered on retest or secondary screening but false negatives represent information that may be lost forever. Bayesian learning has been used to search for, and recover false negatives, increasing the hit rate in follow up screens. The process is as follows:

1. Build a Bayesian model of hit vs baseline from the initial screening results.
2. Rank the inactives from the screen according to the Bayesian model.
3. Since a good model will rank the false negatives towards the top of the list, select a small percentage of the inactive list with the highest Bayesian scores to send for retesting along with the actives.

The protocol below shows a simulation of this process, using the NCI data described above



Half the actives are marked as inactive and the model learned from all the data. The negatives (including the false negatives) are then scored by the model and the rank ordering examined using an enrichment plot.



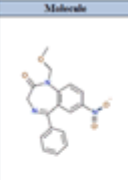
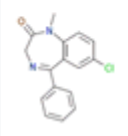
This shows that by retesting the top 5% of the list sorted by the model score, 85% of the false negatives would be recovered.

Bayesian learning has also been successfully applied to iterative screening. Traditional screening tests an entire dataset (e.g. the corporate collection). In iterative screening, only a subset of available data is tested, and the results of this test are used to determine which subset to test next. The procedure is iterated with the aim of ensuring that the earlier rounds of screening will contain most the actives, and potentially mean that screening can be stopped before all iterations are completed. Bayesian modeling is applied after each round of screening to the cumulative results so far, and the remaining candidates to be screened are ranked according to the latest model and the next subset at each round selected from the highest scoring molecules. (REF TO SHEF 2004 IN PRESENTATION ARCHIVE)

Multiple Category Modeling

The score from a Bayesian model is relative. If molecule A scores higher than molecule B in a model, then all that can be said is that A is more likely to have come from the subset of good than B. It doesn't allow any comment to be made about whether A or B are predicted to be good, or how much better A will be than B. The problem extends to two models. If molecule A scores X in model 1 and Y in model 2, where $X > Y$, then nothing can be said about whether A will be more active in screen 1 than screen 2 or not. SciTegic has developed a novel validation procedure to allow an individual model to be calibrated, so that relative scores can be made absolute. With this procedure comparisons can be made, such as those discussed above, to say for example, that molecule A is likely to be active in screen 1 and inactive in screen 2. This has direct application in side effect prediction.

Using multiple category modeling, models of multiple assays (for example the entire corporate collection vs all screens that have been run) can be built and compared. The algorithm may also be applied to drug compendia in a similar way. As an illustration, a model was built for 7500 known drugs from 13 activity classes (including known benzodiazepines). This model was then applied to another small set of benzodiazepines that were not used in the model building, producing the results shown below.

Molecule	Name	SD_Activity/Enrichment	SD_Activity/Enrichment
	BENZODIAZEPINE1	SD_CalciumChannelBlocker	514.8
		SD_Benzodiazepine	197.5
		SD_MAOInhibitor	0.1000
		SD_Sedative	0.1000
		SD_AnticancerDrug	0.1000
		SD_Estrogen	0.000
		SD_EstrogenAntagonist	0.000
		SD_ACEInhibitor	0.000
		SD_Antiproliferative	0.000
		SD_Prostaglandin	0.000
		SD_Corticosteroid	0.000
		SD_BetaBlocker	0.000
		SD_HMGCoAReductaseInhibitor	0.000
	BENZODIAZEPINE2	SD_Benzodiazepine	197.5
		SD_MAOInhibitor	1.300
		SD_Antiproliferative	0.3000
		SD_CalciumChannelBlocker	0.1000
		SD_Sedative	0.1000
		SD_EstrogenAntagonist	0.000
		SD_Estrogen	0.000
		SD_ACEInhibitor	0.000
		SD_AnticancerDrug	0.000
		SD_Prostaglandin	0.000
		SD_Corticosteroid	0.000
		SD_BetaBlocker	0.000
		SD_HMGCoAReductaseInhibitor	0.000

Both candidate molecules are predicted to be benzodiazepines (both 400 times more likely to be than not). However the first molecule is 500 times more likely to be a calcium channel blocker than not, whereas molecule 2 is unlikely to have any of the other activities modeled. As a candidate for further work, molecule 2 therefore looks more interesting than molecule 1.

Copyright © 2001-2008 Accelrys Software Inc.

[Careers](#) | [Legal](#) / [Terms of Use](#) | [Contact us](#)