

# Studying the effect of noise on Laplacian-modified Bayesian Analysis and Tanimoto Similarity

David Rogers, Ph.D.

SciTegic, Inc.

(Division of Accelrys, Inc.)

[drogers@scitegic.com](mailto:drogers@scitegic.com)



# Organization of talk

- Description of:
  - Analysis methods
  - Descriptor
  - Data
- Experimental:
  - ROC Plots of Tanimoto and modified Bayesian
  - Re-run with 10% noise
  - Re-run with 20% noise
- Discussion

## Analysis: Tanimoto similarity

- Tanimoto similarity T
  - Given two fingerprints A and B
  - $T(A, B) = \|\{A\} \text{ and } \{B\}\| / \|\{A\} \text{ or } \{B\}\|$
- Tanimoto prioritization P
  - Given a set of actives/hits  $H_i$
  - Given a test compound C
  - $P(C, H) = \text{MAX}_{i=1, N}(T(H_i, C))$
  - Score is similarity to most-similar hit
- Test set is sorted using P in decreasing order

# Analysis: Laplacian-modified Bayesian

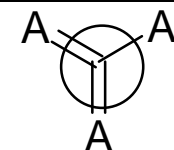
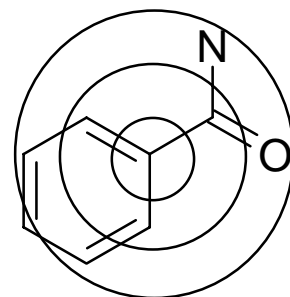
- Probabilistic (vs fitting) method
  - For each feature, keep count:
    - # times it was seen ( $F_{\text{total}}$ )
    - # times it was seen in the actives/hits ( $F_{\text{active}}$ )
  - Estimate probability from counts
  - Combine probabilities from different features
    - (Assuming independence)
- Xiaoyang Xia, Edward G. Maliski, Paul Gallant, and David Rogers, "Classification of Kinase Inhibitors Using a Bayesian Model", *J. Med. Chem.*, 2004, **47**, 4463-4470

## Laplacian-modified Bayesian (II)

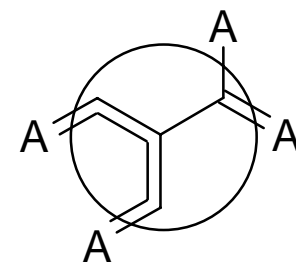
- Not a “pure” Naïve-Bayesian method
  - Effect of *absence* of feature not included
  - Different features are sampled at different rates
- If we had a sufficiently-large number of samples:
  - $P(\text{act}|\text{feature}) = F_{\text{active}} / F_{\text{total}}$ .
- However, if the # of samples is small:
  - Bias from sampling error may be large
- The Laplacian correction can be thought of as adding some number  $N$  of “virtual” samples to the counts, at baseline probability. The corrected estimator is:
  - $P(\text{act}|\text{feature}) = (F_{\text{active}} + N * P(\text{act})) / (F_{\text{total}} + N)$ .

# Descriptor: Extended-connectivity fingerprints

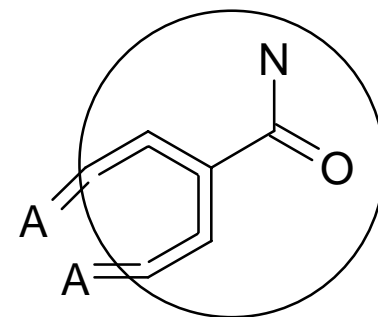
- Derived from the *Morgan algorithm*
  - Morgan H. J. *Chem Doc*, 1965, 5, pp. 107-113
- Initial atom code:
  - H-donor, H-acceptor, positive/negative ionizable, halogen, aromatic
- At each iteration, combine atom code with neighbor's codes and hash



Iteration 0



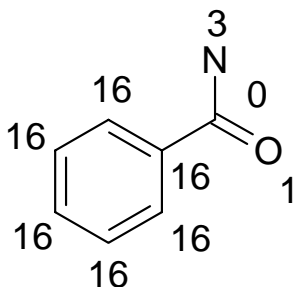
Iteration 1



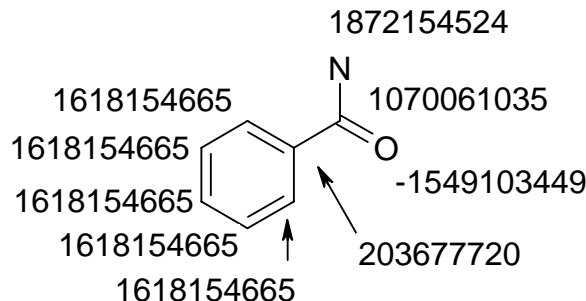
Iteration 2

Each iteration adds bits that represent larger and larger structures

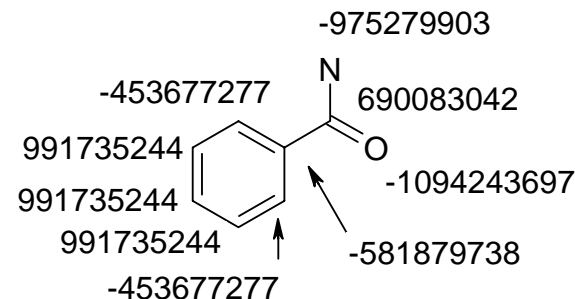
# Extended-connectivity fingerprints (II)



```
> <FCFP_0>
16
0
1
3
```



```
> <FCFP_2>
16
0
1
3
1618154665
203677720
-1549103449
1872154524
1070061035
...
```

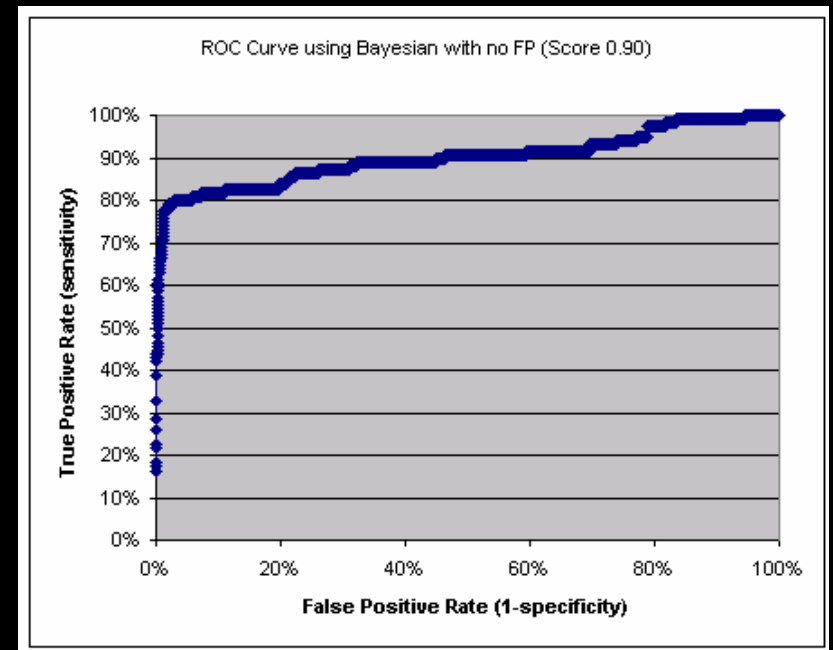
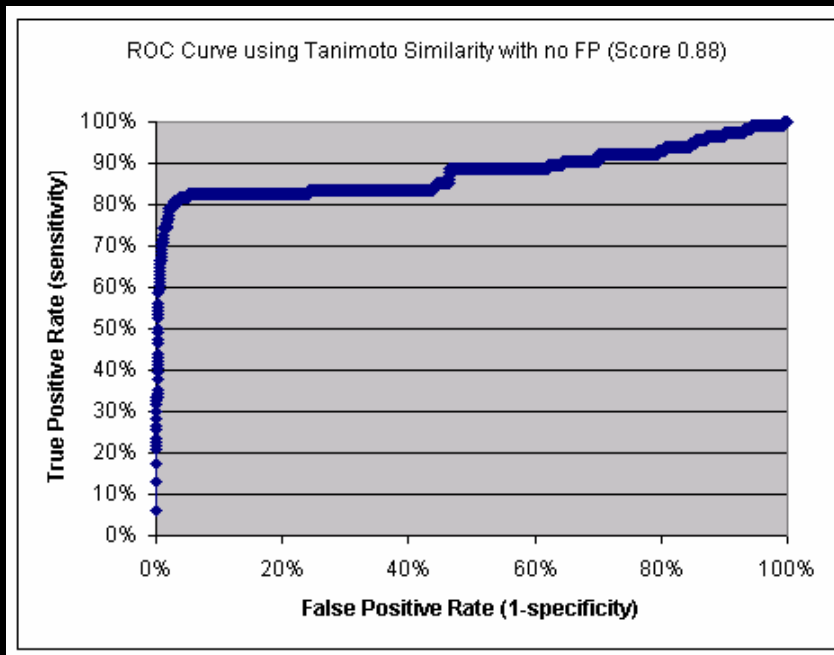


```
> <FCFP_4>
16
0
1
3
1618154665
203677720
-1549103449
1872154524
1070061035
991735244
-453677277
-581879738
-1094243697
690083042
-975279903
```

## The data set

- NCI AIDS data set
  - Inhibition of cell growth of HIV-infected cells
  - ~32,000 compounds
  - 230 confirmed hits
- Data split into training and test sets
  - ~16,000 compounds in each
  - 114/116 true positives in training/test sets
- Noisy training data created by adding *false positives*
  - 0%: 114 TP, 0 FP, 15896 TN
  - 10%: 114 TP, 1594 FP, 14416 TN
  - 20% 114 TP, 3136 FP, 12874 TN

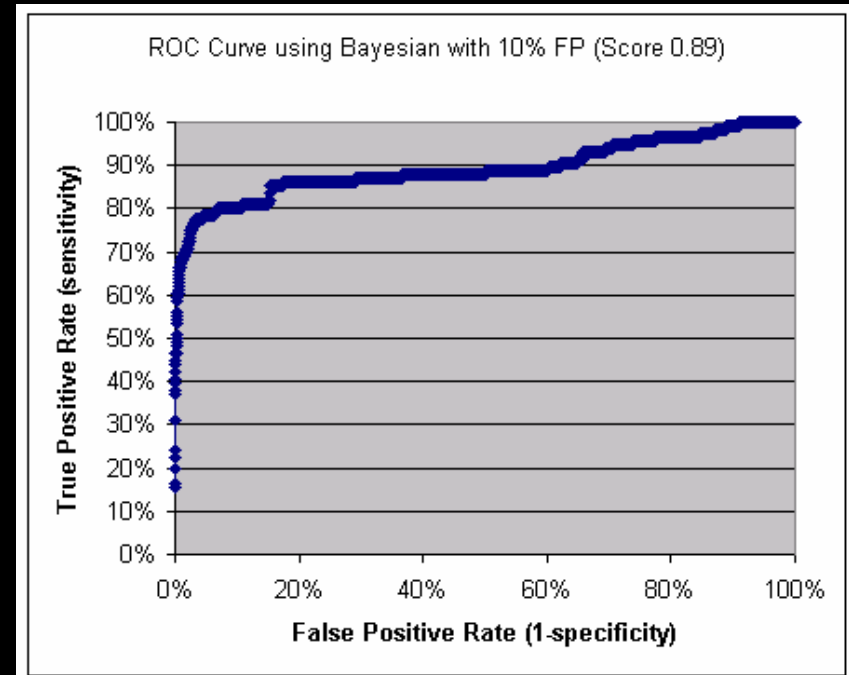
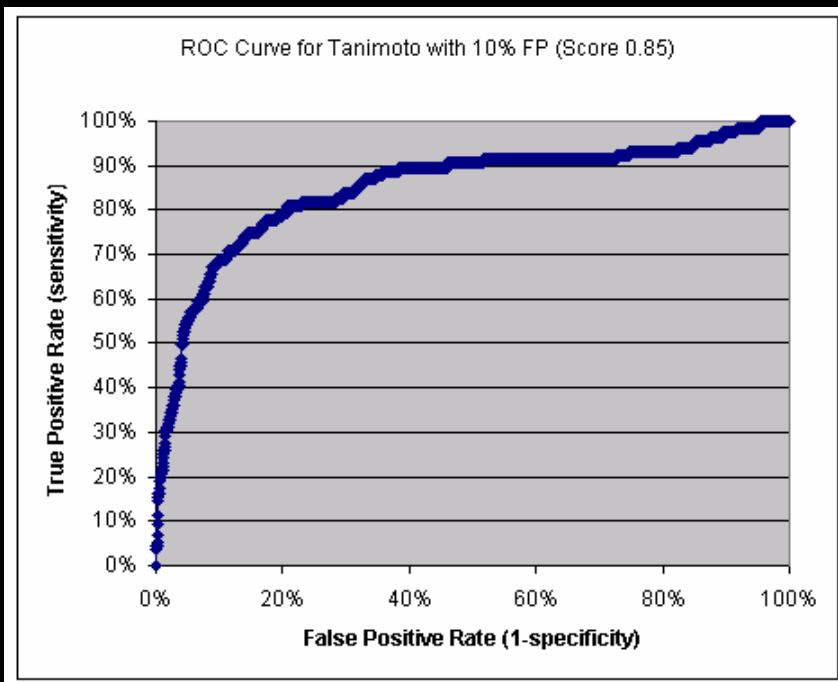
# Experimental: no noise



- Slight benefit to modified Bayesian
- Laplacian adjustment reduces “noise” caused by uninformative features

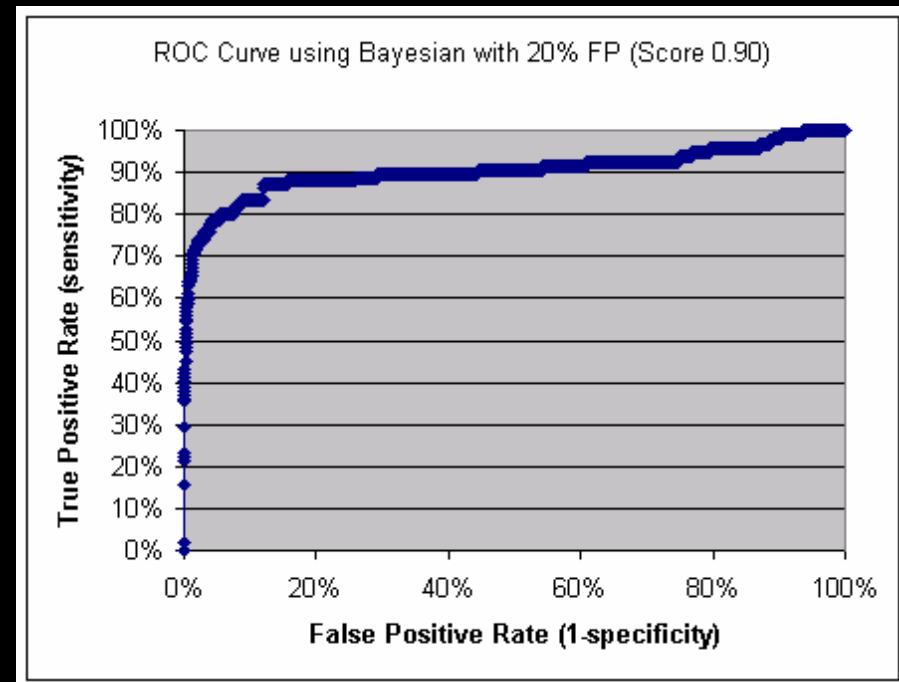
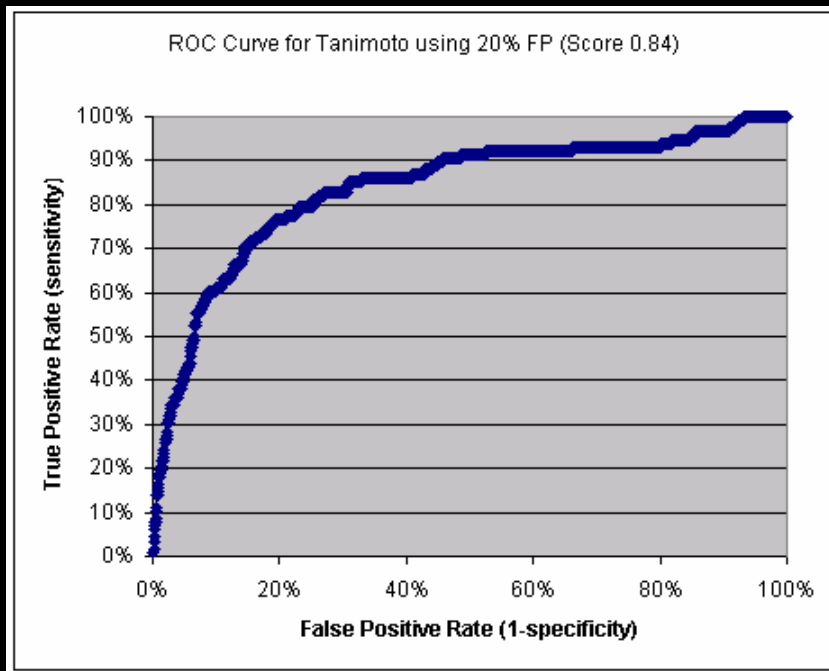
## Experimental: 10% noise

- 1594 true negatives changed to false positives
  - $1594 + 114 = 1708$  “positives”



## Experimental: 20% noise

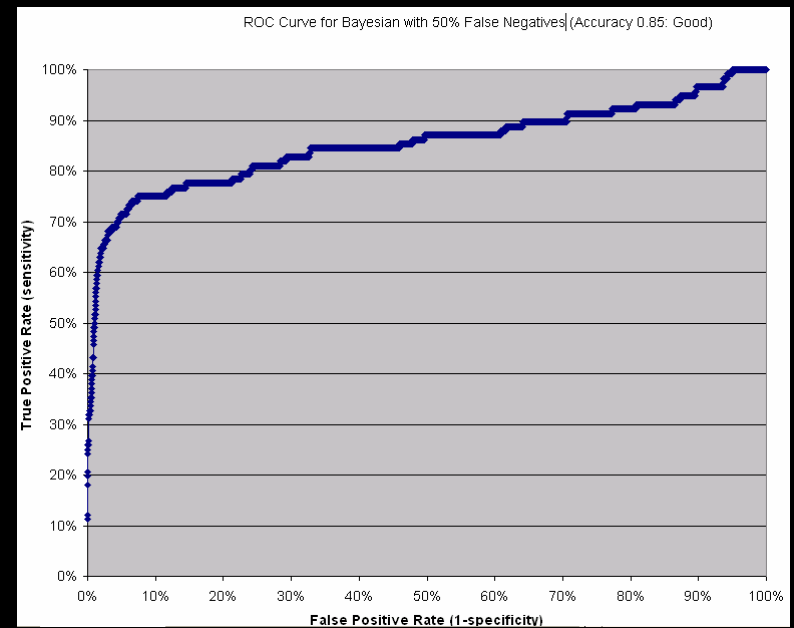
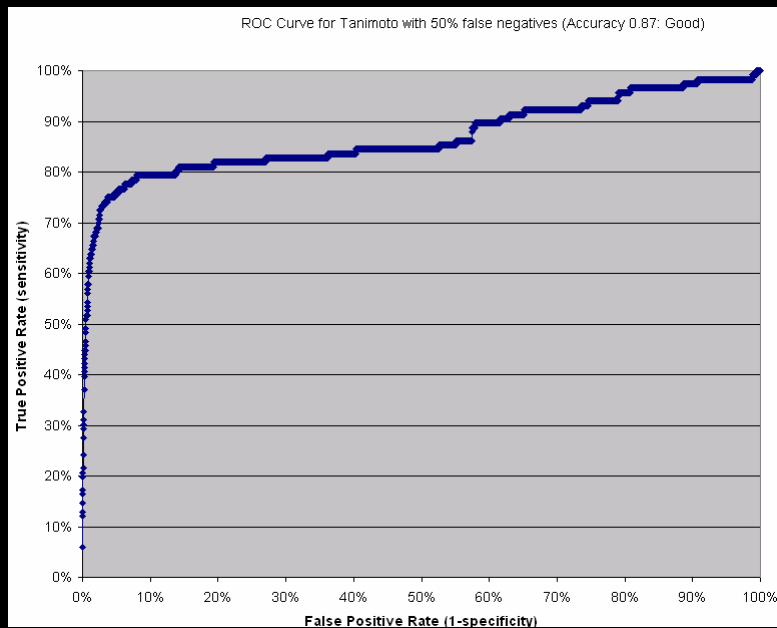
- 3136 true negatives changed to false positives
  - $3136 + 114 = 3250$  “positives”



- Laplacian-modified Bayesian shows little effect

# No free lunch: effect of false negatives

- Took 50% of true positives in the training data and declared them false negatives
  - ROC of .87 for Tanimoto, .85 for modified Bayesian
  - False negative data is *counterweight* to true positive data



# Discussion

- Tanimoto similarity
  - Sensitive to false positives
  - Less sensitive to false negatives (in fact, it ignores them)
  - More expensive to execute with lots of FPs ( $N \times M$  vs.  $N$ )
- Laplacian-modified Bayesian
  - Resists noise caused by uninformative bits
    - Larger problem with larger fingerprints (ECFP\_6, etc.)
  - Resists noise caused by false positives
    - “Snow on the Great Plains”
    - Tanimoto fooled if *any* hit is close
    - Bayesian wants *many* to be
  - Affected by false negatives
    - Things that fail to be found counterweight things that are found

## Thanks to...

- Dr. Andy Liaw (Merck), who suggested this work
- Matt Hahn, early ECFP work

# And now, the bunny...

