The Use of Kernel Discrimination Algorithms in Virtual Screening

David Wood
UK QSAR – April 24th 2007
Overview

• Machine learning and kernel-based algorithms

• Virtual screening experiments with the MDDR

• The effect of noisy training data

• Application to real pharmaceutical HTS datasets

• Summary
Machine Learning

**Training Set**
- Known active compounds
- Known inactive compounds

**Test Set**
- (screening candidates)
  - C1, C2, C3, C4, C5 ...

**Model of Activity**

**Compute scores for Test Set**

**Likelihood of activity**

**Top Ranked Compounds Picked for Screening**
- C3
- C1
- C4
- C2
- C5
- ...
Kernel Discrimination

• A kernel function provides a distance-based weighting for a pair of descriptor vectors

• For binary vectors (fingerprints)...

\[ k(x, y) = \lambda^{M - d_{xy}} (1 - \lambda)^{d_{xy}} \]

• For continuous vectors...

\[ k(x, y) = h^{-1}(2\pi)^{-1/2} e^{-d(x,y)^2/2h^2} \]

• Where:
  • \( M \) is the vector length
  • \( d_{xy} \) is the squared Euclidean distance
  • \( \lambda \) is a smoothing parameter
    Optimised by a Leave-one-out cross validation
    \( 0.5 < \lambda < 1 \)

• Where:
  • \( d(x,y) \) is the Euclidean distance
  • \( h \) is the bandwidth of the Gaussians
    Equivalent to the smoothing parameter
    \( 0 < h \)
Kernel Discrimination

- Given a set of compounds $C$, the probability density at a position in the descriptor space $x$ can be estimated by the sum of the kernel-based weights

$$\hat{p}(x) = \frac{1}{n} \sum_{i \in C} k(x, C_i)$$

- With a set of active $A$ and inactive $I$ compounds, the $KD\_SCORE$ provides the relative probability of activity for an unknown compound $x$

$$KD\_SCORE(x, A) = \frac{\sum_{i \in A} k(x, A_i)}{\sum_{i \in I} k(x, I_i)}$$
Kernel Discrimination: 1-Dimensional Example

- Compounds are distributed throughout the descriptor space.
- The kernel function spreads the ‘mass’ of each example as a Gaussian distribution.
- At any point in the descriptor space, the *probability density* can be estimated as the sum of the Gaussian.
- The width of the Gaussians must be optimised.
  - Leave-one-out cross validation method.
  - Compounds are scored according to the likelihood ratio.
Machine Learning Algorithms

Support Vector Machine

- A relatively new and very effective method
- The SVM projects the descriptor space into a feature space of higher dimensions where classification becomes a linearly separable problem
Machine Learning Algorithms

- Substructural Analysis
  - Otherwise known as a naïve Bayesian classifier
  - Used with binary descriptors, i.e. molecular fingerprints
  - The method identifies substructures that are related to activity or inactivity

\[
\text{R4 weighting scheme score for substructure } j = \log \left( \frac{A_j / N_A}{T_j / N_T} \right)
\]

Where:
- \( A_j \) is the number of occurrences of substructure \( j \) in the set of actives
- \( N_A \) is the number of actives
- \( T_j \) is the number of occurrences of substructure \( j \) in the entire set
- \( N_T \) is the number of compounds
Virtual Screening with the MDDR
Datasets – From the MDDR Database

- MDL Drug Data Report (MDDR) is a database of ~100,000 biologically active compounds collected from:
  - Journals
  - Patent literature
  - Meetings
  - Congress

- 11 activity classes selected from the database

- Five training sets of 100 active and 4000 inactive compounds were generated for each activity class

<table>
<thead>
<tr>
<th>Activity Class</th>
<th>Number of Actives</th>
</tr>
</thead>
<tbody>
<tr>
<td>5HT3 Antagonists</td>
<td>752</td>
</tr>
<tr>
<td>5HT1A Agonists</td>
<td>827</td>
</tr>
<tr>
<td>5HT Reuptake Inhibitors</td>
<td>359</td>
</tr>
<tr>
<td>D2 Antagonists</td>
<td>395</td>
</tr>
<tr>
<td>Renin Inhibitors</td>
<td>1130</td>
</tr>
<tr>
<td>Angiotensin II AT1 Antagonists</td>
<td>943</td>
</tr>
<tr>
<td>Thrombin Inhibitor</td>
<td>803</td>
</tr>
<tr>
<td>Substance P Antagonists</td>
<td>1246</td>
</tr>
<tr>
<td>HIV Protease Inhibitor</td>
<td>750</td>
</tr>
<tr>
<td>Cyclo-oxygenase Inhibitor</td>
<td>636</td>
</tr>
<tr>
<td>Protein Kinase C Inhibitor</td>
<td>453</td>
</tr>
</tbody>
</table>
Descriptors

• Binary vectors:
  • Molecular fingerprints
    Unity, BCI, Daylight and SciTegic’s ECFP4

• Continuous Vectors:
  • Tripos Holograms
    Similar to a molecular fingerprint
    Structural feature counts are encoded into an integer vector
  • Tripos MolconnZ
    A set of approximately 500 topological indices
    Processed with a Principal Components Analysis

• Physicochemical Properties
  A set of 32 properties
  Include MW, LogP, PSA, Number H-bond donors & acceptors
  Processed with a Principal Components Analysis
Comparison of Fingerprints

Training Set = 100 active compounds, 100 inactive compounds
Average of 5 runs

Proportion of actives retrieved in the top 1%

BCI
Daylight
ECFP_4
Unity

5HT3 ANTAG
5HT1A AGON
5HT REUP
INHIB
D2 ANTAG
RENIN INHIB
ANGIO II AT1 ANTAG
THROMB INHIB
SUB P ANTAG
HIV PRO INHIB
COX INHIB
PRO KIN C INHIB
AVERAGE
Comparison of the Kernel Discrimination Algorithm and a Support Vector Machine

Visualising the Models: $5HT_3$ Antagonists
Application to Noisy Training Data
High Throughput Screening Data

• HTS data is widely recognised to be relatively noisy
  • Many false positives
  • Machine learning algorithms must be able to tolerate such noise

• Training sets that mimic noisy HTS data were generated from the MDDR
  • Training sets of 100 active and 4000 inactive compounds represented by SciTegic’s ECFP4 fingerprints were systematically corrupted

• The performance of BKD with noisy input data was compared to that of SSA and SVM
The Effect of Noise in the Training Data: ECFP4 Descriptors

Why is SSA More Tolerant to Noise?

- **SSA**
  - Looks for general structural trends within the training set
  - Is able to ignore structural outliers (false positives)

- **BKD**
  - A localised model – nearest training set neighbours govern the likelihood of activity
  - Cannot disregard structural outliers

- **SVM**
  - Very effective with clean and noisy data
  - Complex mathematics – a black box
Analysis of SSA

The Effect of False Positives on the Magnitude and Sign of the Substructure Weights.

Beneficial Fragments

<table>
<thead>
<tr>
<th>Substructure Weight</th>
<th>0% fp's</th>
<th>40% fp's</th>
<th>80% fp's</th>
<th>95% fp's</th>
</tr>
</thead>
<tbody>
<tr>
<td>Substructure (Bit position)</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

Graph showing the effect of different false positive rates on substructure weights.
Analysis of SSA

The Effect of False Positives on the Magnitude and Sign of the Substructure Weights: A Moving Average.

Beneficial Substructures

<table>
<thead>
<tr>
<th>Local Average Weight</th>
<th>0% fp's</th>
<th>40% fp's</th>
<th>80% fp's</th>
<th>95% fp's</th>
</tr>
</thead>
<tbody>
<tr>
<td>Substructures</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Analysis of a BKD Run: Smoothing Parameter = 0.6

- Influence of the nearest training set neighbours on two highly ranked compounds
  - 80% False Positives

### Highest Ranked Compound

<table>
<thead>
<tr>
<th>Training Set Class</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Training Set Sum</th>
<th>BKD_SCORE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actives</td>
<td>5.52E-01</td>
<td>1.29E-10</td>
<td>7.15E-11</td>
<td>5.86E-11</td>
<td>5.86E-11</td>
<td>5.52E-01</td>
<td>4.04E+07</td>
</tr>
<tr>
<td>Inactives</td>
<td>4.25E-10</td>
<td>1.92E-10</td>
<td>1.29E-10</td>
<td>1.29E-10</td>
<td>1.06E-10</td>
<td>1.37E-08</td>
<td></td>
</tr>
</tbody>
</table>

### 1000th Ranked Compound

<table>
<thead>
<tr>
<th>Training Set Class</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Training Set Sum</th>
<th>BKD_SCORE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actives</td>
<td>4.21E-02</td>
<td>9.58E-07</td>
<td>1.09E-07</td>
<td>8.90E-08</td>
<td>7.30E-08</td>
<td>4.21E-02</td>
<td>5.39E+00</td>
</tr>
<tr>
<td>Inactives</td>
<td>7.09E-03</td>
<td>6.59E-04</td>
<td>4.12E-05</td>
<td>3.83E-06</td>
<td>2.11E-06</td>
<td>7.81E-03</td>
<td></td>
</tr>
</tbody>
</table>
Analysis of a BKD Run: Smoothing Parameter = 0.52

- Influence of the nearest training set neighbours on two highly ranked compounds
  - 80% False Positives

### Highest Ranked Compound

<table>
<thead>
<tr>
<th>Training Set Class</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Training Set Sum</th>
<th>BKD_SCORE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actives</td>
<td>8.55E-01</td>
<td>5.56E-01</td>
<td>4.95E-01</td>
<td>4.40E-01</td>
<td>2.75E-01</td>
<td>3.66E+01</td>
<td>1.78E-01</td>
</tr>
<tr>
<td>Inactives</td>
<td>2.75E-01</td>
<td>2.55E-01</td>
<td>2.35E-01</td>
<td>1.94E-01</td>
<td>1.94E-01</td>
<td>2.06E+02</td>
<td></td>
</tr>
</tbody>
</table>

### 1000th Ranked Compound

<table>
<thead>
<tr>
<th>Training Set Class</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Training Set Sum</th>
<th>BKD_SCORE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actives</td>
<td>2.86E-01</td>
<td>2.75E-01</td>
<td>2.75E-01</td>
<td>2.01E-01</td>
<td>2.01E-01</td>
<td>2.55E+01</td>
<td>1.56E-01</td>
</tr>
<tr>
<td>Inactives</td>
<td>2.55E-01</td>
<td>2.55E-01</td>
<td>2.45E-01</td>
<td>2.45E-01</td>
<td>2.26E-01</td>
<td>1.63E+02</td>
<td></td>
</tr>
</tbody>
</table>
The Effect of the Smoothing Parameter on BKD’s Tolerance to Noise

[Graph showing the effect of smoothing parameter on percentage of actives retrieved in the top 1%]
Application to real pharmaceutical HTS datasets
Real Pharmaceutical HTS Datasets

• The VS methods were applied to a selection of HTS datasets obtained by GSK, represented Daylight fingerprints

  • Training sets:
    1000 HTS hits (actives + false positives)
    5000 HTS non-hits (inactives + false negatives)

  • Test set:
    All remaining compounds

• Scoring the performance:
  The enrichment of the confirmed actives in the test set (true positives – tested for IC50)
HTS Dataset: A 7TM Receptor Target

Receiver Operating Characteristic Curve

<table>
<thead>
<tr>
<th>Area under the ROC curve</th>
<th>Enrichment of actives at 10%</th>
</tr>
</thead>
<tbody>
<tr>
<td>BKD</td>
<td>SVM-POLY</td>
</tr>
<tr>
<td>0.799</td>
<td>0.845</td>
</tr>
</tbody>
</table>
In Summary…

• KD algorithms are very effective when applied to high quality training data

• SVM are more tolerant to noise in the training data than KD

• SSA was found to be extremely tolerant to noise in the training data, but was generally found to be less effective than KD and SVM

• Successful models of activity can be built using pharmaceutical HTS data

• Viewing 1D models of molecular properties can provide insights into the structural preferences of the targets
Acknowledgments

- **Supervisors:**
  - Peter Willett
  - Beining Chen
  - Xiao Lewell
  - Rob Harrison

- **Colleagues:**
  - William Heal
  - Roger Mutter
  - Chido Mpamhanga
  - Jérôme Hert
  - Yogi Patel
  - Kirstin Moffat

- Funding from BBSRC and GlaxoSmithKline
Thank You!