Machine Learning and Drug Design

Observations and Lessons

UCSC Machine Learning Summer School 2012

Nigel Duffy, PhD.
Applying Machine Learning

More than picking the best / hottest algorithm ...

• What is the problem you’re trying to solve?
  • How do you measure improvement or success?
• What data is available?
  • How much?
  • Where does it come from?
  • What does it look like?
• What algorithms best match your problem and your data?

If Weka had all the answers no one would need us ...
An Aside

Classification and Regression Trees, 1984

Has maximum in [0.1,0.34]?
Finding New Drugs

R&D productivity is on the decline

Note: R&D costs are estimated from PhRMA annual survey 2008; NMEs are the total number of small molecule and biologic approvals by the FDA. Source: Bernstein Research "The Long View – R&D Productivity" (September 30, 2010)

Copyright © 2011 by The Boston Consulting Group
Finding New Drugs

What is the learning problem?
Predicting Compound Activity

Thomas Splettstoesser, Wikimedia Commons
Predicting Compound Activity

Encode chemical information – Representation

Capture key aspects of problem

Challenging statistical problems

Overfitting: The “train/test paradox”

Bias: Where does the training data come from?

Noise: What were the experimental conditions?
Representation: Conformers

Enumerate potential binding conformations of small molecules
Representation: 3D Features

• Capture key shape and electrostatic features
Representation: Pharmacophores

Pharmacophore: Set of 4 feature points and distances
  - Rotation and translation invariant
  - Captures local information
  - Doesn’t require whole molecule alignment

Single pharmacophores are bad
  - 4 Points do not capture enough information
  - Not noise tolerant
  - Descriptive but not predictive
Each pharmacophore corresponds to a bit (~10 million bits)
Representation: Sets of Pharmacophores

Bit Keys are bad: Discretized Distances
• Sensitive to conformer generation
• Sensitive to bin boundaries
• Lack resolution
• Poor generalization

Bit Keys are sufficient for purchasable screen
Not sufficient for lead identification

Solution: Use the full set of pharmacophores
Available Data

Literature:
- Curation errors
- Publication bias
- Divergent assay conditions

Patents:
- Curation errors
- Divergent assay conditions
- Obfuscated data
- Only intervals reported
Addressing Noise Using Ranking

- **IC**<sub>50</sub> values are not consistent
  - Experimental noise
  - Varying experimental conditions
- Relative orders are consistent
  - Treat all data as ordering data

Lab 1

Lab 2
IC\textsubscript{50} values are not consistent
  – Experimental noise
  – Varying experimental conditions

Relative orders are consistent
  – Treat all data as ordering data

Tolerance windows address measurement noise

Addressing Noise Using Ranking
## Noise: Some examples

### Varying Assay Conditions:

<table>
<thead>
<tr>
<th>Assay</th>
<th>Compounds</th>
<th>Pearson (log)</th>
<th>Kendall Tau</th>
<th>KT (tol=1.0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HEK v. CHO</td>
<td>425</td>
<td>0.71</td>
<td>0.56</td>
<td>0.92</td>
</tr>
<tr>
<td>HEK Internal</td>
<td>40</td>
<td>0.74</td>
<td>0.64</td>
<td>1.0</td>
</tr>
<tr>
<td>CHO Internal</td>
<td>60</td>
<td>0.96</td>
<td>0.79</td>
<td>1.0</td>
</tr>
</tbody>
</table>
Noise: Some examples

“Same” Conditions – different lab:

<table>
<thead>
<tr>
<th>Assay</th>
<th>Compounds</th>
<th>Pearson (log)</th>
<th>Kendall Tau</th>
<th>KT (tol=0.5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kinase 1</td>
<td>43</td>
<td>0.96</td>
<td>0.78</td>
<td>1.0</td>
</tr>
<tr>
<td>Kinase 2</td>
<td>23</td>
<td>0.91</td>
<td>0.77</td>
<td>1.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Assay</th>
<th>Compounds</th>
<th>Pearson (log)</th>
<th>Kendall Tau</th>
<th>KT (tol=0.5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kinase 3 Assay 1</td>
<td>44</td>
<td>0.82</td>
<td>0.73</td>
<td>0.97</td>
</tr>
<tr>
<td>Kinase 3 Assay 2</td>
<td>45</td>
<td>0.79</td>
<td>0.58</td>
<td>0.99</td>
</tr>
</tbody>
</table>
Bias

Our Goal: Find NEW molecules
Data: Only regarding old molecules – owned by someone else

Publication bias: Mostly active compounds are published

Medicinal chemistry bias: Compounds are made because someone believed they would be active

Synthetic chemistry bias: More easy compounds are made than hard compounds
Empirical Risk Minimization

Minimize the empirical risk:

\[ \sum_{(x,y) \in S} L(f(x), y) \]

- where \( S \) is the training data
- \( f \) is a predictive model
- \( x \) is the representation
- \( y \) is the label
- \( L \) is a bounded loss function

Can alternatively be expressed as

\[ E_{(x,y) \sim S} [L(f(x), y)] \]

Expected loss over the empirical distribution
Bias: Where does our data come from?

Essential assumption: Training data is drawn independently from an *identical* distribution to the test distribution $D$, i.e.,

$$E_{(x,y) \sim S} [L(x, y)] = E_{(x,y) \sim D} [L(x, y)]$$

Due to bias this is *never* even *approximately* true for biochemical data, i.e, $S$ is not equal $D$

Publication bias, Medicinal chemistry bias, Synthetic bias

*Must address data bias for ALL methods*

Including model validation, docking, and free-energy
Bias: Where do we put our errors?

Consider fitting the function $\sin(x/y)$ where

- $X$ (visible) is drawn independently and uniformly from $[-3, 3]$.
- $Y$ (hidden) is 3 or 1 with varying probabilities.
Bias: Where do we put our errors?

What if?

In training data $P(y=1)=0.05$ and $P(y=3)=0.95$
In testing data $P(y=1)=0.95$ and $P(y=3)=0.05$

Makes error estimates worthless $\Rightarrow$ Modeling is hard
Publication Bias

Publication bias → Mostly active compounds are published
Baseline predictor: predict active

Published Data
Apparent active: inactive ratio > 10:1
Predict active: >90% accuracy on literature data

Reality
True active: inactive ratio < 1:1000
Predict active: <0.1% accuracy on real compounds

Where do we put the errors? Not where we should
Addressing Publication Bias with Ranking

Treat all data as ordering data:
- Compound A is active and compound B is inactive
Create ordered pairs of data (A,B)
Are they in the correct order?

Bias disappears, baseline is 50% accurate.
Measuring Success

Our problem:
From a large set of molecules select a small number to make and test

The challenge:
Making and testing is expensive. How do you evaluate methods?

It turns out that the probability of zero hits is:

\[
\left( \frac{an-k}{na} \right)^{1/2} (en-1)
\]

\(a = \) area above ROC curve of model
\(n = \) number of molecules evaluated
\(e = \) proportion of active compounds in original set
\(k = \) number of compounds made and tested

So area above ROC curve is an appropriate metric.
Measuring Success
Metrics aren’t the whole story

Be careful of metrics:
• Molecular weight often predicts well
• Intuitively too good to be true
• Likely a problem with your experiment!
Measuring Success

Evaluation experiments must approximate your problem well:

- Reflect real world biases
  - Cannot just randomly sub-sample your data
- Reflect real world noise
  - Cannot test noise robustness on pristine data
- Pure statistical measures of loss don’t capture everything
  - Remember the molecular weight example

So, identify the right metric but never completely trust it.

You eventually have to test in the real world.
VALIDATING ACTIVITY PREDICTIONS

Built models and screened *in silico* library of commercially available compounds

GPCRs, other receptors, kinases, other enzymes, channels, macromolecular

Assayed diverse predicted active compounds in the laboratory >10 times

<table>
<thead>
<tr>
<th>Representative Examples</th>
<th>Actives</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>p38α MAP Kinase Inhibitor (Ser-Thr Kinase)</td>
<td>31/44 (70%)</td>
<td>6 Cmpds IC$_{50}$ &lt; 1 μM 10 new series In vivo activity</td>
</tr>
<tr>
<td>Acyl-CoA Acyl Transferase Inhibitor (ACAT)</td>
<td>17/25 (68%)</td>
<td>IC$_{50}$ 110 nM</td>
</tr>
<tr>
<td>Transglutaminase 2 Inhibitor (TG2)</td>
<td>13/25 (52%)</td>
<td>Input: 2 series Output: 5 series</td>
</tr>
<tr>
<td>Adrenergic Receptor Antagonist α1a + α1d &gt; α1b</td>
<td>5/8 (62%)</td>
<td>Selectivity ≈ Tamsulosin</td>
</tr>
<tr>
<td></td>
<td>2/8 Selective</td>
<td></td>
</tr>
</tbody>
</table>
Built predictive model for p38α MAPK inhibitors
   Literature IC₅₀ data for 73 compounds
   11 distinct structural classes of inhibitors represented
   No target structural data used
Screened library of 700,000 commercial compounds *in silico*
   44 diverse, predicted active compounds obtained for laboratory testing
31 of 44 compounds (70%) inhibit p38α MAPK by ≥ 50% at 10 µM
   6 compounds with IC₅₀ < 1 µM
   Estimate Enrichment ≥ 700
Actives are structurally diverse => robust scaffold-hopping
   New classes (scaffolds) identified equals number in the training set
p38α MAPK MODEL TRAINING SET

Azoles

Ureas

Quinazolinones

Pyridinopyrimidinones

Pyrrolopyrimidines

Quinazolines

Indoles

Benzophenones

Triazines

Amidobenzamides

Benzenediamides

Copyright Numerate Inc.
STRUCTURALLY DIVERSE HITS

- 31 / 44 compounds tested inhibit enzyme activity by ≥ 50% at 10 μm
- Actives include 10 new structural classes
Goal: Protein kinase R inhibitor for TB, diabetes

Collaboration with Carl Nathan (Cornell)
Starting point: 20 PKR inhibitors in 5 series (BBRC 2003 308:50)

Step 1: Augment training data with 28 Ser/Thr Ki’s
Step 2: Identify diverse new PKR inhibitor chemotypes
Build model, screen 5M purchasables in silico
Acquire and assay 162 diverse, predicted actives

35/162 IC\textsubscript{50} < 20 \textmu M, 8 new series
(BMCL 2011 21:4108)

NMRT-2862 is non-toxic to macrophages and recapitulates effects of PKR knockout.
Goal: *Transglutaminase 2 inhibitor for CS,CF*

Collaboration with Chaitan Khosla (Stanford)

Starting point: 35 (μM+) actives in 2 series
Different binding sites/MOA

**Step 1:** Build 1st generation model, screen purchasables
4 new active series identified (A-D)

**Step 2:** Hit expansion in Series D
50 analogs, NIH-funded (*BMCL* 2011 21:2692)

<table>
<thead>
<tr>
<th>Compound</th>
<th>Series</th>
<th>Ki (μM)</th>
<th>MOA</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMRT-1002</td>
<td>A</td>
<td>3</td>
<td>Rev</td>
</tr>
<tr>
<td>NMRT-2149</td>
<td>B</td>
<td>6.5</td>
<td>Rev</td>
</tr>
<tr>
<td>NMRT-2235</td>
<td>C</td>
<td>5.9</td>
<td>Irrev</td>
</tr>
<tr>
<td>NMRT-2081</td>
<td>D</td>
<td>0.4</td>
<td>Rev, Non-comp</td>
</tr>
</tbody>
</table>

Copyright Numerate Inc
Goal: Versatile NNRTI for treatment of HIV/AIDS

Profile

Potent inhibitor of WT and drug-resistant HIV-1 replication
Improved PK/ADME profile and pharmaceutical properties

Models built with whole-cell (phenotypic) data

Huge chemical space outlined

Broad-spectrum lead series achieved in 3 months

<table>
<thead>
<tr>
<th>Compound</th>
<th>WT</th>
<th>L100I</th>
<th>K103N</th>
<th>Y181C</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUSTIVA</td>
<td>1.4 nM</td>
<td>23 nM</td>
<td>12 nM</td>
<td>1.9 nM</td>
</tr>
<tr>
<td>#1</td>
<td>3.8</td>
<td>9.1</td>
<td>13</td>
<td>200</td>
</tr>
<tr>
<td>#21</td>
<td>4.0</td>
<td>6.6</td>
<td>3.2</td>
<td>27</td>
</tr>
</tbody>
</table>
Goal: **Multi-targeted cardioprotective agent**

**Therapeutic profile**
- Potent reductions in cholesterol, triglycerides (TGs), C-reactive protein (CRP), fibrinogen (FN), glucose
- Enhanced anti-inflammatory and cardioprotective properties
- Low potential for myopathy and drug-drug interactions

**Pharmacologic profile**
- Potent inhibitor of HMG-CoA reductase
- Potent inhibitor of p38α mitogen-activated protein kinase
  - ↓ TNFα, IL-1β, CRP, FN, TGs, platelet reactivity, atherogenesis
  - ↓ Gluconeogenesis, leptin resistance

- Low affinity for retinoid X receptor α
- Enhanced anti-inflammatory activity
In vitro results

13/19 compounds inhibit HMG-CoA R, 10/19 inhibit p38α MAPK
Dual-acting profile verified in 8/19 compounds
Little or no affinity for RXRα
Improved anti-inflammatory activity in whole cell assay
U.S. Patent claims issued for 3 series

<table>
<thead>
<tr>
<th>Compound</th>
<th>HMG CoA R</th>
<th>p38α MAPK</th>
<th>RXRα</th>
<th>TNFα</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lipitor</td>
<td>7 nM</td>
<td>&gt; 100000 nM</td>
<td>38000 nM</td>
<td>93000 nM</td>
</tr>
<tr>
<td>Baycol</td>
<td>1.3</td>
<td>&gt; 100000</td>
<td>8100</td>
<td>82000</td>
</tr>
<tr>
<td>MTCP-1</td>
<td>33</td>
<td>11000</td>
<td>&gt; 100000</td>
<td>31000</td>
</tr>
<tr>
<td>MTCP-2</td>
<td>2.0</td>
<td>36000</td>
<td>&gt; 100000</td>
<td>30000</td>
</tr>
<tr>
<td>MTCP-3</td>
<td>3.7</td>
<td>11000</td>
<td>80000</td>
<td>21000</td>
</tr>
<tr>
<td>MTCP-4</td>
<td>130</td>
<td>1900</td>
<td>&gt; 100000</td>
<td>18000</td>
</tr>
<tr>
<td>MTCP-5</td>
<td>11</td>
<td>6000</td>
<td>&gt; 100000</td>
<td>230000</td>
</tr>
<tr>
<td>MTCP-6</td>
<td>1.2</td>
<td>400</td>
<td>&gt; 100000</td>
<td>8000</td>
</tr>
<tr>
<td>MTCP-7</td>
<td>4.8</td>
<td>270</td>
<td>&gt; 100000</td>
<td>9700</td>
</tr>
<tr>
<td>MTCP-8</td>
<td>98</td>
<td>&lt; 100</td>
<td>&gt; 100000</td>
<td>9700</td>
</tr>
</tbody>
</table>
In vivo results

ApoE*3 Leiden transgenic mouse model
Significant Cholesterol reduction
Triglyceride reduction ≥ atorvastatin
Reduction in Fibrinogen & E-Selectin ≥ atorvastatin
Conclusions

Successful applications of machine learning require thought:
• Define the problem
• It’s not just about the “best” algorithm
• Match the algorithm to the problem
• Represent the key properties of the data
• Consider the bias
• Think about the noise
• Design your evaluation experiments carefully
• Validate in the real world.