Using Information from Historical High-Throughput Screens to Predict Active Compounds

Sereina Riniker
10th ICCS and 10th GCC, Noordwijkerhout, 2014
Outline

- Introduction
  - High-throughput screening
  - Hit expansion
  - HTS fingerprints

- Learning with HTS fingerprints

- Validation with in-house assays

- Validation with public assays
High-Throughput Screening (HTS)

- HTS today:
  - Many advances in miniaturization, automation, management systems
  - Testing of 1-5 million compounds in a few weeks

- Over 120 full-deck (> 1 million) screens in past 20 years at Novartis

- Limitations on the size of the compound library to be screened:
  - Assay throughput
  - Reagent cost
  - Availability of reagents (cells or proteins)

- Current trend:
  - Use of smaller but smarter sublibraries
  - Diversity selection (chemical and/or biological)

Problem:
- When only a (diversity-selected) sublibrary is screened, what about the actives in the rest of the deck?

Standard in-house procedure to date:
- PipelinePilot Bayes model with ECFP4 fingerprints trained on the molecules in the subset
- Generate predictions for the molecules in the rest of the deck → propose some for the validation screen

Questions:
- Can we do better?
- Can the wealth of information contained in historical HTS assays be used for hit expansion?

HTS Fingerprints

Basic idea

- Chemical biological descriptors based on historical HTS data
  - Describes bioactivity profile of a compound

- Developed by Petrone et al. (2012) using 195 Novartis assays
  - Biochemical and cellular assays
  - Vectors of Z-scores calculated from percent inhibition values

- Similarity metric: SimScore
  - Combination of:
    - Pearson correlation coefficient
    - Number of assays in common
  - Gaps in the fingerprint are taken into account

- Applications:
  - Similarity search and scaffold hopping
  - Subset design: plate diversity selection


Learning with HTS Fingerprints

Current set-up: in-house assays

- HTS fingerprints for ~1.7 million compounds:
  - To reduce the number of gaps, focus on full-deck screens: 93 full-deck primary HTS screens
  - Bits = normal Z-scores
    \[ Z(x_i) = x_i - \frac{\text{mean}(x)}{\text{std}(x)} \]
  - Gaps: assume to be mean Z-score (i.e. zero)

- Test sets:
  - Set A: pair of in-house assays: 250K (for training) and full-deck (for testing)
  - Set B: 50 in-house assays with 750K - 1.2 million compounds

- Machine-learning (ML) methods:
  - Random forest (RF), logistic regression (LR), naïve Bayes (NB)
Results

In-house test set A

- Comparison RF(HTS-fp) to PipelinePilot (PP) NB(ECFP4)

<table>
<thead>
<tr>
<th>Method</th>
<th>AUC</th>
<th>EF(0.1%)</th>
<th>EF(0.5%)</th>
<th>EF(1%)</th>
<th>EF(5%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PP-NB(ECFP4)</td>
<td>0.838</td>
<td>22.53</td>
<td>25.16</td>
<td>20.41</td>
<td>9.10</td>
</tr>
<tr>
<td>scikit-learn RF(HTS-fp)</td>
<td>0.864</td>
<td>51.60</td>
<td>35.87</td>
<td>28.41</td>
<td>11.25</td>
</tr>
</tbody>
</table>

Why is RF(HTS-fp) so good?
→ Is it related targets or frequent hitters? Or, are we just geniuses?

EF(0.1%) = 51.6 means:
719 of top 1049 molecules are primary actives (69%)
Possible Bias: Related Targets?

*In-house test set A*

- What do the HTS fingerprints of the top 200 actives look like?

- What are the targets of these five assays?
  - Oxidoreductase, protein-tyrosine kinase, \(N\)-methyltransferase and phosphatase
  - None closely related to the glycosyltransferase studied in current assay
Possible Bias: Frequent Hitters?

*In-house test set A*

- Likelihood that an HTS-fp with many high/low Z-scores will be at the top of the ranked list

![Graph showing the number of assays where a molecule is active vs. rank]

- What is the effect of frequent hitters?
  - Remove molecules active in more than 10 assays from training/testing

<table>
<thead>
<tr>
<th>Method</th>
<th>AUC</th>
<th>EF(0.1%)</th>
<th>EF(0.5%)</th>
<th>EF(1%)</th>
<th>EF(5%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF(HTS-fp)</td>
<td>0.864</td>
<td>51.60</td>
<td>35.87</td>
<td>28.41</td>
<td>11.25</td>
</tr>
<tr>
<td>RF(HTS-fp) – no freq. hits</td>
<td>0.855</td>
<td>46.82</td>
<td>33.99</td>
<td>26.45</td>
<td>10.87</td>
</tr>
</tbody>
</table>
Possible Biases

In-house test set A

- Large bias by frequent hitters and/or related targets can be excluded

- Baseline model:
  
  - Rank compounds by median(Z-scores)
    - If the actives of an assay have neg. Z-scores → take the inverted values

<table>
<thead>
<tr>
<th>Method</th>
<th>AUC</th>
<th>EF(0.1%)</th>
<th>EF(0.5%)</th>
<th>EF(1%)</th>
<th>EF(5%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF(HTS-fp)</td>
<td>0.864</td>
<td>51.60</td>
<td>35.87</td>
<td>28.41</td>
<td>11.25</td>
</tr>
<tr>
<td>median(Z-scores)</td>
<td>0.769</td>
<td>20.60</td>
<td>13.23</td>
<td>11.70</td>
<td>7.33</td>
</tr>
</tbody>
</table>

- Much lower performance, but also not random!
Combination with Chemical Similarity

*In-house test set A*

- **Best "chemical" model:**
  - Logistic regression (LR) classifier with unfolded (dictionary-based) Morgan2 fingerprints (Morgan2 = RDKit implementation of ECFP4)

- **Taking advantage of both descriptions:**
  - Classifier fusion of RF(HTS-fp) and LR(Morgan2)

<table>
<thead>
<tr>
<th>Method</th>
<th>AUC</th>
<th>EF(0.1%)</th>
<th>EF(0.5%)</th>
<th>EF(1%)</th>
<th>EF(5%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF(HTS-fp)</td>
<td>0.864</td>
<td>51.60</td>
<td>35.87</td>
<td>28.41</td>
<td>11.25</td>
</tr>
<tr>
<td>LR(Morgan2)</td>
<td>0.860</td>
<td>54.83</td>
<td>37.43</td>
<td>28.24</td>
<td>10.75</td>
</tr>
<tr>
<td>Classifier fusion</td>
<td><strong>0.903</strong></td>
<td><strong>55.19</strong></td>
<td><strong>41.42</strong></td>
<td><strong>31.81</strong></td>
<td><strong>12.75</strong></td>
</tr>
</tbody>
</table>

→ Classifier fusion is working.

- **In addition, solution to a limitation of RF(HTS-fp):**
  - Including the chemical model allows molecules without "HTS-history" to be considered
In-house test set A

**How good are the models at recognizing true positives?**

- Total 16'265 mols tested in validation assay $\rightarrow$ 4748 validated hits

<table>
<thead>
<tr>
<th>Method</th>
<th>Cutoff</th>
<th>N_{act}</th>
<th>N_{val}</th>
<th>ratio [%]</th>
<th>N_{mol}</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF(HTS-fp)</td>
<td>0.1</td>
<td>355</td>
<td>537</td>
<td>66.1</td>
<td>1049</td>
</tr>
<tr>
<td>LR(Morgan2)</td>
<td>0.1</td>
<td>271</td>
<td>337</td>
<td>80.4</td>
<td>1049</td>
</tr>
<tr>
<td>Classifier fusion</td>
<td>0.1</td>
<td>359</td>
<td>475</td>
<td>75.6</td>
<td>1049</td>
</tr>
</tbody>
</table>

**RF(HTS-fp) finds most actives (absolute number)**

<table>
<thead>
<tr>
<th>Method</th>
<th>Cutoff</th>
<th>N_{act}</th>
<th>N_{val}</th>
<th>ratio [%]</th>
<th>N_{mol}</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF(HTS-fp)</td>
<td>1.0</td>
<td>1184</td>
<td>2401</td>
<td>49.3</td>
<td>10499</td>
</tr>
<tr>
<td>LR(Morgan2)</td>
<td>1.0</td>
<td>899</td>
<td>1536</td>
<td>58.5</td>
<td>10499</td>
</tr>
<tr>
<td>Classifier fusion</td>
<td>1.0</td>
<td>1418</td>
<td>2474</td>
<td>57.3</td>
<td>10499</td>
</tr>
</tbody>
</table>

**And if the models are trained with only validated actives?**

<table>
<thead>
<tr>
<th>Method</th>
<th>Cutoff</th>
<th>N_{act}</th>
<th>N_{val}</th>
<th>ratio [%]</th>
<th>N_{mol}</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF(HTS-fp)</td>
<td>0.1</td>
<td>368</td>
<td>744</td>
<td>49.5</td>
<td>1049</td>
</tr>
<tr>
<td>LR(Morgan2)</td>
<td>0.1</td>
<td>268</td>
<td>782</td>
<td>34.3</td>
<td>1049</td>
</tr>
<tr>
<td>Classifier fusion</td>
<td>0.1</td>
<td>347</td>
<td>788</td>
<td>44.0</td>
<td>1049</td>
</tr>
</tbody>
</table>

**using validated hits for training improves quality of the model**
Prospective Test

In-house test set A

- Opportunity to propose ~2K compounds for validation based on fusion approach
  - Older version of approach: RF(binary HTS-fp) + NB(ECFP4) (using PipelinePilot)
  - Selection: top 2188 molecules not already tested (i.e. primary inactives)
  - 2015 of the 2188 compounds were available for testing
  - **1230 of the 2015 compounds were found active (61%)**

- How many of these were top ranked by current approach?

<table>
<thead>
<tr>
<th>Method</th>
<th>Cutoff</th>
<th>N_{act}(ML)</th>
<th>N_{val}(ML)</th>
<th>ratio [%]</th>
<th>N_{mol}(ML)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF(HTS-fp)</td>
<td>0.1</td>
<td>194</td>
<td>225</td>
<td>86.2</td>
<td>512</td>
</tr>
<tr>
<td>LR(Morgan2)</td>
<td>0.1</td>
<td>143</td>
<td>257</td>
<td>55.6</td>
<td>712</td>
</tr>
<tr>
<td><strong>Classifier fusion</strong></td>
<td><strong>0.1</strong></td>
<td><strong>211</strong></td>
<td><strong>302</strong></td>
<td><strong>69.9</strong></td>
<td><strong>574</strong></td>
</tr>
<tr>
<td>RF(HTS-fp)</td>
<td>1.0</td>
<td>865</td>
<td>1099</td>
<td>78.7</td>
<td>8098</td>
</tr>
<tr>
<td>LR(Morgan2)</td>
<td>1.0</td>
<td>332</td>
<td>674</td>
<td>49.3</td>
<td>8963</td>
</tr>
<tr>
<td><strong>Classifier fusion</strong></td>
<td><strong>1.0</strong></td>
<td><strong>876</strong></td>
<td><strong>1287</strong></td>
<td><strong>68.1</strong></td>
<td><strong>8025</strong></td>
</tr>
</tbody>
</table>
The results presented before were for a single assay (pair)

In-house test set B:
- 50 in-house (biochemical and cellular) assays with 750K – 1.2 million compounds
- 10 repetitions with randomly selected 200K training molecules
- Targets: broad set of kinases, proteases, ion channels, GPCRs and other target classes, but different from the targets of the historical assays

How do the performances look like across a broad range of assays?
Results

In-house test set B

Average performance over 10 repetitions.

Per repetition: randomly selected 200K training set, the rest for testing

Fusion is consistently best over a broad range of assays.
In order to investigate and present the approach in more details (which is not possible with the in-house set), we repeated the process with assays from PubChem.

**Same procedure:**

- **HTS fingerprints** for ~430K compounds:
  - 95 assays from PubChem with > 338K compounds

- **Test sets**:
  - 46 assays from PubChem with 300K – 338K compounds
  - 10 repetitions with randomly selected 50K training molecules
Results

Public test set (PubChem)

Fusion is consistently best over a broad range of assays.

Average performance over 10 repetitions.
Per repetition: randomly selected 50K training set, the rest for testing.

<table>
<thead>
<tr>
<th>ICCS 2014</th>
<th>Sereina Riniker</th>
<th>June 1-5, 2014</th>
</tr>
</thead>
</table>

- RF(HTS-fp)
- LR(Morgan2)
- Classifier fusion
- Median(Z-scores)

AUC

EF(5%)
**Results**

*Public test set (PubChem)*

- Scaffold hopping potential of HTS fingerprints

Morgan3 similarity between actives in top 5% and training actives

- HTS fingerprints retrieve more diverse actives
- Fusion is in between HTS-fp and Morgan2
Results

Public test set (PubChem)

- Comparison of BMS enrichment in top 5% and EF(5%)

\[
\begin{align*}
\text{Classifier fusion} & : r^2 = 0.9771 \\
\text{LR(Morgan2)} & : r^2 = 0.9670 \\
\text{RF(HTS-fp)} & : r^2 = 0.9774
\end{align*}
\]
Summary

- Learning with HTS fingerprints is a promising approach for hit expansion.

- Combination of biological and chemical information through classifier fusion of RF(HTS-fp) and LR(Morgan2) provides robust performance across a broad range of assays.

- HTS fingerprints have higher scaffold-hopping potential than Morgan2.

- Approach works for both in-house and public HTS data.
Acknowledgements

- Greg Landrum
- Yuan Wang, Jeremy Jenkins
- Nikolas Fechner
- Allen Cornett, Maxim Popov, Ansgar Schuffenhauser
- Novartis Education Office