Effective Analysis of Data Mining Results

Marvin Waldman and Osman F. Güner
Molecular Simulations Inc.
San Diego, CA

A Subsidiary of Pharmacopeia, Inc.
Overview

• Introduction/background - analysis of hit lists
• Metrics for analyzing hit lists
• GH-Score: Performance and recent developments
• Review of recent applications
• Application to virtual HTS
• Future directions
Database Domain

Schematic representation of a typical database and a hit list that contains some known active compounds.
Analyzing Hit Lists

Different metrics can be used to evaluate the quality of a hit list

- **Enrichment** ($E$): indicates the ratio of yield of actives in the hit list relative to the yield of actives in the database.

$$ E = \left( \frac{H_a / H_t}{A / D} \right) = \frac{H_a \times D}{H_t \times A} $$

where:
- $H_t$ is the total number of compounds in the hit list
- $H_a$ is the number of known actives in the hit list
- $A$ is the number of active compounds in the hit list
- $D$ is the number of compounds in the database.
Analyzing Hit Lists

- **Yield (\(\%Y\))** is the percentage of known actives in the hit list

\[
\%Y = \frac{H_a}{H_t} \times 100
\]

- **Coverage (\(\%A\))** is the percentage of known active compounds retrieved from the database

\[
\%A = \frac{H_a}{A} \times 100
\]
GH-Score

\[
GH = w \left( \frac{H_a}{H_t} \right) + (1 - w) \left( \frac{H_a}{A} \right)
\]

\[
GH = \frac{H_a [wA + (1 - w)H_t]}{AH_t}
\]

with \( w = 0.75 \), and correction for database size

\[
GH = \left( \frac{H_a (3A + H_t)}{4H_t A} \right) \times \left( 1 - \frac{H_t - H_a}{D - A} \right)
\]

Testing the Metrics Against the Best and Worst Hit Lists

- **The “Best” Hit List**
  - $A = H_a = H_t$

- **The “Worst” Hit List**
  - $H_t = D - A$
  - $H_a = 0$
Metrics Perform Consistently Against Typical Cases

<table>
<thead>
<tr>
<th>Case</th>
<th>Enrichment</th>
<th>%Y</th>
<th>%A</th>
<th>GH-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best</td>
<td>500</td>
<td>100</td>
<td>100</td>
<td>1</td>
</tr>
<tr>
<td>Good</td>
<td>200</td>
<td>40</td>
<td>80</td>
<td>0.5</td>
</tr>
<tr>
<td>Bad</td>
<td>25</td>
<td>5</td>
<td>50</td>
<td>0.16</td>
</tr>
<tr>
<td>Worst</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

where $D = 50,000$, $A = 100$

**Good** hit list is 80 active compounds in a hit list of 200

**Bad** hit list is 50 active compounds in a hit list of 1,000
### Against Extreme Cases?

<table>
<thead>
<tr>
<th>Case</th>
<th>Enrichment</th>
<th>%Y</th>
<th>%A</th>
<th>GH-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best</td>
<td>500</td>
<td>100</td>
<td>100</td>
<td>1</td>
</tr>
<tr>
<td><strong>ExtremeY</strong></td>
<td><strong>500</strong></td>
<td><strong>100</strong></td>
<td>1</td>
<td>0.75</td>
</tr>
<tr>
<td>Good</td>
<td>200</td>
<td>40</td>
<td>80</td>
<td>0.5</td>
</tr>
<tr>
<td>Bad</td>
<td>25</td>
<td>5</td>
<td>50</td>
<td>0.16</td>
</tr>
<tr>
<td><strong>ExtremeA</strong></td>
<td>1</td>
<td>0.2</td>
<td><strong>100</strong></td>
<td>0</td>
</tr>
<tr>
<td>Worst</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

where \( D = 50,000 \), \( A = 100 \)

**Good** hit list is 80 active compounds in a hit list of 200

**Bad** hit list is 50 active compounds in a hit list of 1,000

**ExtremeY** is a single hit that is active

**ExtremeA** retrieves the entire database
## Effects of weights and Database correction term on GH-score

<table>
<thead>
<tr>
<th>Case</th>
<th>Equal wts</th>
<th>Uneq wts</th>
<th>Eq+DB</th>
<th>Uneq+DB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ExtremeY</td>
<td>0.51</td>
<td>0.75</td>
<td>0.51</td>
<td>0.75</td>
</tr>
<tr>
<td>Good</td>
<td>0.6</td>
<td>0.5</td>
<td>0.6</td>
<td>0.5</td>
</tr>
<tr>
<td>Bad</td>
<td>0.28</td>
<td>0.16</td>
<td>0.27</td>
<td>0.16</td>
</tr>
<tr>
<td>ExtremeA</td>
<td>0.5</td>
<td>0.25</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Worst</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

where $D = 50,000$, $A = 100$

**Good** hit list is 80 active compounds in a hit list of 200

**Bad** hit list is 50 active compounds in a hit list of 1,000

**ExtremeY** is a single hit that is active

**ExtremeA** retrieves the entire database

---

Copyright © 2001 Marvin Waldman and Osman Guner

A Subsidiary of Pharmacopeia, Inc.
### 3D-Search Query Analysis

<table>
<thead>
<tr>
<th>Query</th>
<th># Actives (Ha)</th>
<th># Hits (Ht)</th>
<th>%Y</th>
<th>%A</th>
<th>Enrichment (E)</th>
<th>GH score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Database</td>
<td>80</td>
<td>10,318</td>
<td>0.78</td>
<td>100.0</td>
<td>1.0</td>
<td>0</td>
</tr>
<tr>
<td>Shape</td>
<td>13</td>
<td>2,244</td>
<td>0.58</td>
<td>16.3</td>
<td>0.8</td>
<td>0.035</td>
</tr>
<tr>
<td>Pharmacophore</td>
<td>23</td>
<td>1,144</td>
<td>2.01</td>
<td>28.8</td>
<td>2.6</td>
<td>0.077</td>
</tr>
<tr>
<td>Merged</td>
<td>4</td>
<td>20</td>
<td><strong>20.00</strong></td>
<td>5.0</td>
<td><strong>25.8</strong></td>
<td><strong>0.163</strong></td>
</tr>
</tbody>
</table>

Other Published Uses of GH-Score

• Similarity Search Retrieval Effectiveness

• Query Optimization

• Virtual High-throughput Screening
Virtual HTS: Docking and Scoring

See:

J. Med. Chem. 2000, 43, 4759-4767

“Protein-Based Virtual Screening of Chemical Databases. 1. Evaluation of Different Docking/Scoring Combinations”
Preparation of the Virtual Screening Data Set and the Target Protein

- 9 known TK inhibitors
  - random conformation
  - energy minimized

- 993 randomly selected structures from WDI database
  - molecular weight < 500
  - keep largest fragment
  - energy minimized

- PDB complex: pdb1kim
  - Delete ligand and water molecules
  - Add hydrogens from templates (Cerius²)

Total: 1002 molecules
Binding site of 1kim based on protein shape (default)

- Overlay of 1kim ligand and the default binding site
Docking Protocol: Cerius² LigandFit

- Force field: Dreiding
- Dielectric constant: 4\epsilon
- Softened vdw potential
- Modified trilinear interpolation on energy grid
- Number of Monte Carlo torsion trials: 5000
- Number of saved fits: 10
- Rigid body minimization

**Total docked structures: 2402**
Scoring protocols

• LigScore - Dreiding

• PLP (Piecewise Linear Potential)
  Gehlhaar, D.K. et al. Rational drug design (ACS symposium series 719)
  292-311 (1999)

• PMF (Potentials of Mean Force)

• Ludi
True Hit Rate of TK Inhibitors Using Different Scoring Functions

- Ligscore
- PLP v.1
- PMF
- LUDI_1

Graph showing True Hit Rate on the Y-axis and Rank on the X-axis.
Consensus scoring of TK inhibitors by Ligscore and PLP v.1
Result of consensus scoring using Ligscore and PLP

Top 10% Ligscore
100 Molecules with 9 known inhibitors

Top 10% PLP v.1
100 Molecules with 7 known inhibitors

Consensus score of Ligscore + PLP v.1
34 Molecules with 7 known inhibitors

GH=0.288
GH=0.224
GH=0.339
Future Directions

• Use GH Score as objective function to automatically train/optimize a query

• Refinements to GH Score with additional criteria
  • Worst Hit List: GH = -1  
    (Result is worse than random selection)
  • Enrichment = 1 ⇒ GH = 0  
    (Result is no better than random selection)
Acknowledgements

• C. M. Venkatachalam
• Jeff Jiang

Thymidine Kinase Study