From QSAR to Big Data: Developing Mechanism-Driven Predictive Models for Animal Toxicity

Hao Zhu

Department of Chemistry
The Rutgers Center for Computational and Integrative Biology
Rutgers University-Camden
Email: hao.zhu99@rutgers.edu

September 24, 2015
Acknowledgements

Rutgers:
• PhD students: Marlene Kim, Wenyi Wang, Daniel Russo, Linlin Zhao
• Master students: Kathryn Ribay, Joe Hess
• Visiting Scholar: Dr. Aleck Sedykh, Dr. Jun Zhang

John Hopkins University:
• Dr. Thomas Hartung

Shandong University:
• Dr. Bing Yan

NCATS:
• Dr. Menghang Xia, Dr. Ruili Huang

ICCVAM:
• Dr. Judy Strickland

Funding resource:
• National Institute of Health: 1R15ES023148
• Society of Toxicology: Colgate-Palmolive Grant for Alternative Research
Toxicity evaluation today

Principles of QSAR modeling

Quantitative Structure Activity Relationships

Slide Courtesy of Dr. Fourches
Principles of QSAR modeling

With m molecules and n descriptors

\[ f \begin{pmatrix} \text{Pattern matrix} \\ i \end{pmatrix} = \text{PROPERTY} (i) \]
The “similar” compounds that have “dissimilar” toxicity profiles

<table>
<thead>
<tr>
<th></th>
<th>IN VIVO ASSAYS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MOUSE_KIDNEY</td>
</tr>
<tr>
<td></td>
<td>RAT_SKELETAL_AXIAL</td>
</tr>
<tr>
<td></td>
<td>MGR_RAT_LIVER</td>
</tr>
<tr>
<td></td>
<td>MGR_RAT_KIDNEY</td>
</tr>
<tr>
<td>METOLACHLOR</td>
<td>0 0 0 0</td>
</tr>
<tr>
<td>NN1-ACETOCHLOR</td>
<td>1 1 1 1</td>
</tr>
<tr>
<td>NN3-METALAXYL</td>
<td>- 0 - -</td>
</tr>
</tbody>
</table>

**NOTES:**
0: non-toxic/inactive
1: toxic/active
-: not tested
NN: Nearest Neighbor
PubChem data in 2014

- >700,000 bioassays
- >200,000,000 bioactivity outcomes
- >1,200,000,000 data points
- >2,800,000 small molecule samples
- >1,900,000 chemical structures
- >108,000 RNAi reagents

Chemical-in vitro-in vivo profiles in big data era

Chem. Res. Tox. 2014; (27) 1643-1651
Before the ToxCast project, data already existed

Obtained from PubChem on Aug. 1, 2013, before the ToxCast phase II data was released.
The current question is:

- What can we do if we have limited in-house data available for the compounds of interest?
Antioxidant Response Element $\beta$-lactamase reporter gene assay (ARE-bla)

• Recognized by the Tox21 program as one of the most important toxicity assays
• ARE genes play a role in alleviating oxidative stress

Reactive Oxygen Species (ROS)

Oxygen
\( \text{O}_2 \)

Superoxide anion
\( \cdot \text{O}_2^- \)

Peroxide
\( \cdot \text{O}_2^- \)

Hydrogen Peroxide
\( \text{H}_2\text{O}_2 \)

Hydroxyl radical
\( \cdot \text{OH} \)

Hydroxyl ion
\( \text{OH}^- \)

Liver damage
Workflow for profiling liver toxicants

1. BIG DATA
   - Liver Damage
   - ~1,300 compounds

2. Relevant Assays
   - qHTS ARE-bla
   - ~10,000 compounds

3. Mechanism Analysis & Compound Prioritization
   - ARE
   - IIC

<table>
<thead>
<tr>
<th>Compound</th>
<th>Liver Tox</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>...</th>
<th>n</th>
<th>ARE</th>
<th>IIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>FN</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>FP</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>TP</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>TN</td>
</tr>
<tr>
<td>5</td>
<td>Untested compounds</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>Non-toxic</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>Non-toxic</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>Toxic</td>
</tr>
</tbody>
</table>
Profiling target compounds with biological responses using automated tool

• Input – target compounds:
  1. qHTS ARE-bla dataset (10,928 compounds)
  2. FDA liver damage dataset (1,314 compounds)

• Output – assays related to:
  1. qHTS ARE-bla activation (1,819 assays)
  2. Liver damage (1,159 assays)
Criteria for filtering inadequate and finding relevant assays

- 2,978 Initial number of assays retrieved
- 958 Must appear in both groups (qHTS ARE-bla and liver damage)
- 20 Contained >10 true positive responses
- 14 Correlation was better than random (CCR >50%)
- 14 In vitro assay
- 4 Evidence supported by reliable literature
Individual assays showed poor IIC, but the combined response using RA>0.25 show statistical significance.

<table>
<thead>
<tr>
<th>Bioassay</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liver damage activity</td>
<td>--</td>
</tr>
<tr>
<td>686978</td>
<td>0.59</td>
</tr>
<tr>
<td>743067</td>
<td>0.42</td>
</tr>
<tr>
<td>743140</td>
<td>0.24</td>
</tr>
<tr>
<td>(ARE-bla) 743202</td>
<td>0.31</td>
</tr>
<tr>
<td>Combined</td>
<td>4.25x10^{-4}</td>
</tr>
</tbody>
</table>

Active or toxic
Inconclusive or untested
Inactive or non-toxic

Rate of actives (RA)

\[ RA = \frac{A}{A + I} \]

A = no. of active responses
I = no. in active responses
Modeling qHTS ARE-\textit{bla} activation using QSAR approaches: 5-fold cross validation for all individual models
Evaluating In vitro-In vivo Correlations (IICs)

- Focused on compounds that were active in qHTS ARE-bla and liver toxic
- Searched for common chemical features
- Evaluated IICs (sensitivity, specificity, CCR, and $\chi^2$)
IIC between qHTS ARE-\textit{bla} activation and liver damage for overlapping compounds containing the toxicophores

\begin{itemize}
  \item A) ARE-bla
    \begin{itemize}
      \item CCR = 0.64
      \item QSAR CCR = 0.60
    \end{itemize}
  \item B) ARE-bla
    \begin{itemize}
      \item CCR = 0.61
      \item QSAR CCR = 0.57
    \end{itemize}
\end{itemize}
3-D plot of Tox21 phase II modeling set vs FDA liver damage dataset using principal components analysis
Liver toxicity mechanism analysis involving ARE pathway perturbations

**Stimulus**
- Electrophilic fragments, reactive oxygen species, stress

**Nrf2**
- ARE 
- PPAR-γ

**TDP1**
- DNA repair
- (AID 686978)

**Oxidative Stress**
- Thyroid damage
- Liver damage

**TR**
- (AID 743067)
Conclusions

• Developed a workflow
  – Profiles biological responses from big data
  – Incorporates QSAR models to fill-in missing data
  – Evaluates the chemical IIC

• Identified toxicophores and assays that can be used to assess liver damage induced by oxidative stress

• Workflow can be adapted to model or assess other complex animal toxicity endpoints

Take home message

• Reliable information exists, but it is difficult to locate
• Good data may not guarantee good decisions