Use Of *In Silico Tools* for Toxicity Prediction Within ICH M7

Alex Cayley
Senior Scientist
alex.cayley@lhasalimited.org
Overview

• Introduction to the ICH M7 guideline
• Different approaches to toxicity prediction
• Performance of predictive systems
• Improving upon the raw performance metrics
### Table 1: Impurities Classification with Respect to Mutagenic and Carcinogenic Potential and Resulting Control Actions

<table>
<thead>
<tr>
<th>Class</th>
<th>Definition</th>
<th>Proposed action for control (details in Section 7 and 8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Known mutagenic carcinogens</td>
<td>Control at or below compound-specific acceptable limit</td>
</tr>
<tr>
<td>2</td>
<td>Known mutagens with unknown carcinogenic potential (bacterial mutagenicity positive*, no rodent carcinogenic data)</td>
<td>Control at or below acceptable limits (appropriate TTC)</td>
</tr>
<tr>
<td>3</td>
<td>Alerting structure, unrelated to the structure of the drug substance; no mutagenicity data</td>
<td>Control at or below acceptable limits (appropriate TTC) or conduct bacterial mutagenicity assay; If non-mutagenic = Class 5 If mutagenic = Class 2</td>
</tr>
<tr>
<td>4</td>
<td>Alerting structure, same alert in drug substance or compounds related to the drug substance (e.g., process intermediates) which have been tested and are non-mutagenic</td>
<td>Treat as non-mutagenic impurity</td>
</tr>
<tr>
<td>5</td>
<td>No structural alerts, or alerting structure with sufficient data to demonstrate lack</td>
<td>Treat as non-mutagenic impurity</td>
</tr>
</tbody>
</table>

If warranted, the outcome of any computer system-based analysis can be reviewed with the use of expert knowledge in order to provide additional supportive evidence on relevance of any positive, negative, conflicting or inconclusive prediction and provide a rationale to support the final conclusion.

A computational toxicology assessment should be performed using (Q)SAR methodologies that predict the outcome of a bacterial mutagenicity assay (Ref. 6). Two (Q)SAR prediction methodologies that complement each other should be applied. One methodology should be expert rule-based and the second methodology should be statistical-based. (Q)SAR models utilizing these prediction methodologies should follow...
ICH M7 and Lhasa software

ICH Harmonised Tripartite Guideline

Assessment and Control of DNA Reactive (Mutagenic) Impurities in Pharmaceuticals to Limit Potential Carcinogenic Risk

ICH M7

- Database searching
- Expert prediction
- Statistical prediction
- Expert assessment
- Test

Impurity identification

Purge mitigation

Classification

Archiving

Reporting

Control
Origins of expert mutagenicity prediction

Chemical structure, Salmonella mutagenicity and extent of carcinogenicity as indicators of genotoxic carcinogenesis among 222 chemicals tested in rodents by the U.S. NCI/NTP.

Ashby, L. & Tennant, R.W.

C(X)₆₄
X = H, F, Cl, Br, I (in any combination)
Development of expert mutagenicity predictions

132 alerts for Mutagenicity in Derek Nexus 2018.1

Ridings et al.; Toxicology; 1996; 106; 267-279

1996 2000 2010 2019

58 alerts 60 alerts 91 alerts

specificity  sensitivity  DfW DX

Negative predictions

1996 2000 2010
Origins of statistical mutagenicity predictions

Experimental Data

Fragmentation

Fragment Dictionary

Fragment Association

Predict

Match Fragments

MCase

TOPKAT

Sarah

TOPKAT

Query

Compound

Hanser et al.; J. Cheminform.; 2014; 6; 21
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4048587/
SOHN methodology

Recursive-partitioning leading to Hypotheses

Hierarchical Network

Self Organisation

Self organising hypothesis networks: a new approach for representing and structuring SAR knowledge
Hanser et al; J. Cheminform.; 2014; 6; 21
SOHN methodology

Query structure

matches
SOHN methodology

Query structure
SOHN methodology

Query structure

matches

matches
SOHN methodology

Query structure

matches

matches

matches

matches

matches

matches

matches

matches

matches

matches
Development of statistical mutagenicity predictions

For the Mutagenicity in vitro endpoint the prediction is: POSITIVE with 59% confidence.
A confidence in a prediction can be generated for both expert and statistical approaches.

This measure should reflect accuracy in prediction.

Can be used to identify predictions requiring further analysis.

Can be used to identify situations where a prediction cannot be made.
Confidence in Sarah Nexus

Is the confidence in the accuracy of a prediction which considers:

**Likelihood (certainty of the result)** and **Reliability (level of trust of the model)**

Concordance of supporting data
Concordance of multiple hypotheses

Relevance of supporting data
Quality and Quantity of data

[Diagram showing the relationships between Likelihood and Reliability with different combinations of Low/Low, High/Low, Low/High, and High/High, with labels for Weak Negative, Weak Positive, Strong Positive, and Moderate Positive.]
Confidence in Sarah Nexus

Accuracy = \frac{TP+TN}{TP+TN+FP+FN}

Where
- TP = true positives
- TN = true negatives
- FP = false positives
- FN = false negatives

Validation dataset: Vitic Intermediates (721 molecules)

See also:
Confidence in Derek Nexus

- Reasoning levels used in Derek Nexus to assign confidence

- Negative predictions assign confidence in a negative result

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Derek Reasoning Level</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Positive</strong></td>
<td>Certain</td>
</tr>
<tr>
<td></td>
<td>Probable</td>
</tr>
<tr>
<td></td>
<td>Plausible</td>
</tr>
<tr>
<td></td>
<td>Equivocal</td>
</tr>
<tr>
<td><strong>Negative</strong></td>
<td>Inactive with misclassified and unclassified features</td>
</tr>
<tr>
<td></td>
<td>Inactive with misclassified features</td>
</tr>
<tr>
<td></td>
<td>Inactive with unclassified features</td>
</tr>
<tr>
<td></td>
<td>Inactive</td>
</tr>
<tr>
<td></td>
<td>Improbable</td>
</tr>
</tbody>
</table>

**Misclassified Feature**
Shares a feature with Ames positive compounds which have not activated an alert

**Unclassified Feature**
Contains a feature which is not present in Lhasa mutagenicity database


[Judson et al.](https://pubs.rsc.org/en/content/articlehtml/2013/tx/c2tx20037f); Toxicology research; 2013; 2; 70-79
Combining confidences

### ICH M7 Summary Results

2 predictions related to ICH M7 (for Mutagenicity in Bacterium) have been run for this structure.

<table>
<thead>
<tr>
<th>Type</th>
<th>Endpoint</th>
<th>Species</th>
<th>Result</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICH M7 Prediction</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Derek</td>
<td>Mutagenicity in vitro</td>
<td>bacterium</td>
<td>INACTIVE</td>
<td>Derek KB 2018 1.1</td>
</tr>
<tr>
<td>Sarah</td>
<td>Mutagenicity in vitro</td>
<td>bacterium</td>
<td>NEGATIVE (46%)</td>
<td>Sarah Model - 2.0</td>
</tr>
</tbody>
</table>

---

#### Prediction Breakdown

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Derek Reasoning Level</th>
<th>Number of Bins</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>Certain</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Probable</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Plausible</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Equivocal</td>
<td>1</td>
</tr>
<tr>
<td>Negative</td>
<td>Inactive with misclassified and unclassified features</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Inactive with misclassified features</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Inactive with unclassified features</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Inactive</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Improbable</td>
<td>4</td>
</tr>
</tbody>
</table>

#### Sarah Confidence Level Breakdown

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Sarah Confidence Level</th>
<th>Number of Bins</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>100%</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>66% - 99%</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>33% - 65%</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>0% - 32%</td>
<td>1</td>
</tr>
<tr>
<td>Negative</td>
<td>0% - 32%</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>33% - 65%</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>67% - 99%</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>100%</td>
<td>4</td>
</tr>
<tr>
<td>Equivocal</td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>
Applicability domain

Is based on presence of unknown fragments

for each fragment in the query
  if fragment in training set
  then there is coverage for this fragment

if part of the structure remains not covered
then the query is outside the domain
Physicochemical properties can also be used as descriptors by both types of model (CLogP, HOMO, LUMO)

<table>
<thead>
<tr>
<th>Expert Rule</th>
<th>Statistical Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlation usually causative</td>
<td>Correlation may not be causative</td>
</tr>
<tr>
<td>Slower to implement</td>
<td>Quick to implement</td>
</tr>
<tr>
<td>Rules can be based on theory alone</td>
<td>Large data set required</td>
</tr>
<tr>
<td>Highly interpretable</td>
<td>May not be as interpretable</td>
</tr>
<tr>
<td>Able to deal with ‘noise’ in the data</td>
<td>More prone to errors in data</td>
</tr>
<tr>
<td>Risk of overfitting</td>
<td>Risk of overfitting</td>
</tr>
</tbody>
</table>

Barber et al.; Regul. Toxicol. Pharmacol.; 2017; 84; 124-130
How well do they perform?

- Perform well against Ames mutagenicity data sets but will depend on chemical space

5 different (Q)SARs
Statistical and expert rule based approaches

<table>
<thead>
<tr>
<th>Performance Metric</th>
<th>Average Performance (Public)</th>
<th>Average Performance (Proprietary)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Balanced Accuracy</td>
<td>77%</td>
<td>66%</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>74%</td>
<td>58%</td>
</tr>
<tr>
<td>Specificity</td>
<td>81%</td>
<td>73%</td>
</tr>
<tr>
<td>Coverage</td>
<td>95%</td>
<td>92%</td>
</tr>
</tbody>
</table>

Barber et al.; Reg. Tox. Pharm.; 2016; 76; 7-20
How well do they perform?

- Expert review improves results
- Consensus improves results


Consensus predictions

- **9%**: Likely to conclude positive. Very strong evidence would be needed to overturn both predictions.
- **5%**: Likely to conclude positive. Lack of a second prediction suggests insufficient evidence to draw any other conclusion.
- **15%**: Uncertain. Likely to conclude positive without strong evidence to overturn a positive prediction.

**In silico prediction 1**
- Positive
- Positive
- Positive
- Negative
- Negative

**In silico prediction 2**
- Positive
- O.O.D. or equivocal
- Negative
- O.O.D. or equivocal
- Negative

**16%**: Uncertain. Conservatively could assign as positive. May conclude negative with strong evidence showing feature driving a ‘no prediction’ is present in the same context in known negative examples (without deactivating features).

**55%**: Likely to conclude negative. Expert review should support this conclusion — e.g. by assessing any concerning features (misclassified, unclassified, potentially reactive...)

O.O.D. = out of domain
Consensus predictions

- Results against NIHS QSAR challenge data (12,000 compounds, 86% negative)
Consensus predictions

- Results against NIHS QSAR challenge data (12,000 compounds, 86% negative)
Consensus predictions

- Results against NIHS QSAR challenge data (12,000 compounds, 86% negative)
Expert review

<table>
<thead>
<tr>
<th>Structural assessment</th>
<th>Predicted Ames result</th>
<th>Number of compounds(^a)</th>
<th>Number of Ames positive(^b) (%)</th>
<th>Number of Ames negative(^c) (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Q)SAR</td>
<td>Negative</td>
<td>566</td>
<td>35 (6)</td>
<td>531 (94)</td>
</tr>
<tr>
<td>(Q)SAR + expert eval.</td>
<td>Negative</td>
<td>408</td>
<td>5 (1)</td>
<td>403 (99)</td>
</tr>
</tbody>
</table>

\(^a\) Total number of compounds predicted negative.
\(^b\) The number of Ames positive predictions.
\(^c\) The number of Ames negative predictions.

Expert review

- Expert review requires a number of different criteria to be fulfilled
  - Ability to view the evidence used to build a prediction
  - Knowledge of chemistry and biology associated with the endpoint in order to judge adequacy of data used to make prediction
  - Knowledge of strengths and limitations of predictive approaches in order to judge adequacy of interpretation of the evidence being made
Transparency in predictions

Basis of prediction?
Evidence supporting hypothesis?
Detail of evidence?

Chemical reactivity
Functional groups
Mechanisms of activity

In silico prediction system

Prediction
Confidence
Hypotheses
Supporting examples
Compound metadata
Primary data sources

Prediction results
Positive
30%

H₂N-Aryl

+ve in TA98 with S9
Mutagenesis 2018

Cayley et al.; Mutagenesis; 2019; 34; 25-32
Transparency in predictions

- Functional groups
- Chemical reactivity
- Supporting data
- Similarity
- Interpretation of strain data
- Reactive metabolites
- Protocol and limitations of Ames assay
- Mechanisms of activity
Expert review

Predictions Agree
High Confidence
Relevant Hypotheses
Relevant NN
Reliable Data

Predictions Agree
Lower Confidence
Relevant Hypotheses
Less relevant NN
Reliable Data
Other, more relevant NN available

Predictions Disagree
Low Confidence
NN Not Relevant
Predictions can be resolved by considering common Limitation + additional data

Other, more relevant NN available

Software supplies all required
Evidence can be found
May not resolve
Choosing an *In Silico* System

- Conforms to the OECD principles
  - Defined endpoint
  - Unambiguous algorithm
  - Applicability domain
  - Performance
  - Mechanism
- Updated as new knowledge emerges
- Predicts well in the area of chemical space in which you are interested
- Supplies enough information to allow for expert review
Conclusions

• ICH M7 guidelines present a practical approach to impurity assessment

• *In silico* predictions play a prominent role in the workflow

• Two different complementary methods of *in silico* prediction are required and improve results

• Expert review of results obtained will improve predictivity and accuracy of conclusions

• Prediction systems should supply enough detail to facilitate expert review

• Expert review will vary but should be straightforward in most cases

Obrigado!
Questions?