The Future of Predictive Toxicology in a Flattening World:

Barriers to greater validation and acceptance of QSAR models in regulatory applications

Ann Richard
National Center for Computational Toxicology
US Environmental Protection Agency
Toxicity
Prediction
Problem

Toxicity
Risk Assessment

Problem

in vivo response - disease
in vitro response - bioassays
biofunctional response
protein chemical interactions
chemical structure & property
receptor binding
DNA damage
cell-cell communication
cancer
public policy
neurotox
developmental tox
reprotox
free radicals
P-450 metabolism
mutagenicity
cytotoxicity
acute tox
bioavailability
exposure
economics

Toxicity
Prediction
Problem
Toxicity Risk Assessment

Increasing complexity
Increasing relevance to RA
Increasing uncertainty

SAR
Structure-Activity Relationships

Problem
Prediction
Toxicity
OECD Principles \textit{(Setubal, 2002)}: 

A (Q)SAR model for regulatory purposes should be associated with the following information:

- a defined endpoint
- an unambiguous algorithm, i.e. transparency
- a defined domain of applicability
- appropriate measures of goodness-of-fit, robustness, and predictivity
- a mechanistic interpretation, if possible

\textit{A (Q)SAR prediction for regulatory purposes should be placed in a larger context of information to assess the utility of the prediction.}
Mode of Action: “Use of mode of action information in hazard characterization ... is a central part of the proposed approach ...”

Structural Analogue Data: “Confidence in conclusions is a function of how similar the analogues are to the agent under review in structure, metabolism, and biological activity.”

Structural-Activity Relationships: “SAR analyses and models can be used to predict molecular properties, surrogate biological endpoints, and carcinogenicity. ... Suitable SAR analysis [of chemicals that do not bind covalently to DNA] requires knowledge or postulation of the probable mode(s) of action ...”

Narrative: “biological plausibility”, “mode of action consistent with generally agreed-upon principles and understanding of carcinogenicity”
Important SAR Modeling Concepts:

- **Validation:**
  *What is the performance of model outside of the training set? in different chemical domains?*

- **Domain(s) of applicability:**
  *When should a model be applied? When not?*

- **Confidence level:**
  *What confidence can one place in the prediction?*

*Knowledge includes knowing what you don’t know*
Global SAR Toxicity Prediction Model

Q

Chemical Structures

Descriptors

Activities

Toxicity prediction

Confidence in prediction

low

high
Global SAR Toxicity Prediction Model

Chemical Structures
- Model assumptions
- Training database
- Data quality
- Statistical performance measures

Descriptors
- Prospective predictivity
- Model failures
- Model complexity
- Interpretability

Activities
- Activity prediction
- Consistent with known mechanism of action
- Biological & chemical plausibility
- Model performance for analogues
- Identified analogues
- Relevant descriptors
- Coverage
- Statistical certainty

Confidence in prediction
- high
- low
Example: DEMETRA Project

Hybrid System for prediction of aquatic toxicity
(combination of 5 Neural Network models)

- What sorts of chemicals are well predicted?
- poorly predicted?

For a test chemical prediction:
- What are closest model analogs?
- How accurate are predictions for the closest *model* analogs?
- How well are the closest *chemical* analogs predicted?
- Are descriptors well within range of dataset?
- Is prediction plausible?

**TRAINING SET**

\[ y = 0.90x + 0.26 \]

\[ R^2 = 0.76 \]

**TEST SET**

\[ y = x + 0.52 \]

\[ R^2 = 0.72 \]
Skin Sensitization: Reaction Mechanistic Applicability Domains for Structure–Activity Relationships

Aynur O. Aptula,* Grace Patlewicz,‡ and David W. Roberts§

The prediction of skin sensitization potential with minimum animal testing is currently of great importance in light of forthcoming legislation. A number of structure–activity relationships for skin sensitization have been published over the years, but their applicability has often been limited to structural classes. The concept of an applicability domain for a quantitative structure–activity relationship [(Q)SAR] is increasingly being viewed as key for the predictive application of (Q)SARs. This is particularly the case for skin sensitization if more widely applicable SARs are to be developed. In this paper, we analyze a recently published chemical data set for skin sensitization, apply reaction mechanistic criteria to domain classification, and evaluate the structure–activity trends observed within each of these mechanistic domains.

Table 1. Structural and Reaction Mechanistic Classification of the Gerberick et al. (8) Data Set

<table>
<thead>
<tr>
<th>structural classa</th>
<th>no. of chemicals</th>
<th>reaction mechanistic class</th>
<th>no. of chemicals</th>
</tr>
</thead>
<tbody>
<tr>
<td>aliphatic</td>
<td>9</td>
<td>Michael acceptors and pro-Michael acceptors</td>
<td>12</td>
</tr>
<tr>
<td>aldehydes</td>
<td>3</td>
<td>S_N,Ar electrophiles</td>
<td>1</td>
</tr>
<tr>
<td>ketones</td>
<td>4</td>
<td>S_N,2 electrophiles</td>
<td>4</td>
</tr>
<tr>
<td>halogenated</td>
<td>2</td>
<td>Schiff base formers</td>
<td>7</td>
</tr>
<tr>
<td>compounds</td>
<td></td>
<td>acylating agents</td>
<td>3</td>
</tr>
<tr>
<td>aromatic esters</td>
<td></td>
<td>no evident electrophilic or pro-electrophilic features</td>
<td>14</td>
</tr>
<tr>
<td>aliphatic amines</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>aliphatic alcohols</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>miscellaneousb</td>
<td>11</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Draize Rabbit Eye Test Compatibility with Eye Irritation Thresholds in Humans: A Quantitative Structure-Activity Relationship Analysis

Michael H. Abraham,*† Mostafa Hassanisadi,*‡ Mehdi Jalali-Heravi,† Taravat Ghafourian,*‡ William. S. Cain,§ and J. Enrique Cometto-Muñiz§

**Draize rabbit eye test scores:**
68 Bulk liquids – direct eye exposure
Modified maximum average scores (MMAS)

**Adjusted by liquid-saturated vapor pressure, \( P^0 \):**

\[
\log \left( \frac{1}{EIT} \right) = \log \left( \frac{MMAS}{P^0} \right)
\]

**Human eye irritation thresholds:**
23 compounds - vapor phase exposure
Eye Irritation Threshold (EIT)

**Single QSAR for combined set of 91 compounds:**
5 solubility-related parameters, \( R^2 = 0.936 \)

- Supports relevance of Draize test to humans
- Mechanistic basis of QSAR: passive transfer
- Applicable to low vp chemicals
- Applicable to hazardous chemicals
- Extends range of both datasets
Combining SAR and Biofunctional Information

Predicting Carcinogenicity of Organophosphates

Organophosphate

- In vitro genotoxicity?
- Rapid hydrolysis?
- High acute toxicity?
- Increase serum alkaline phosphatase?

SAR: Structure suggestive of alkylation?

- Strong chelating property?

SAR

- In vivo genotoxicity?

LM

Mar

M

HM

CHEMICAL X
Prioritized for weight of evidence
Determination of
Genotoxic Carcinogenicity

Available Empirical Data
On Carcinogenicity

Sufficient Weight of Evidence
for/against Carcinogenicity?

YES

Available Empirical Data
On Genotoxicity

Sufficient Weight of Evidence
for/against Genotoxicity?

NO

YES

SAR Model Predictions
For Carcinogenicity

SAR Model Predictions
For Genotoxicity

QSAR Model Predictions
For Carcinogenicity

QSAR Model Predictions
For Genotoxicity

Sufficient Weight of Evidence –
Positive or Negative - for
Carcinogenicity?

WEIGHT OF EVIDENCE
CONCLUSION FOR
CARCINOGENICITY

NO

YES

WEIGHT OF EVIDENCE
CONCLUSION FOR
GENOTOXIC CARCINOGENICITY

Model Robustness
Indicators

Coutesy of Bette Meek, Health Canada
Integrated Testing Scheme: New Chemical Hazard Assessment

Physchem Properties
- MP, BP, solub, LogP, VP, ...

Tox Properties
- acute oral
- subacute tox (28 day)
- acute dermal/inhalation
- eye/skin irritation
- skin sensitization
- in vitro mutag (2 tests)

Ecotox Properties
- biodeg, acute daphnia, ...
- acute fish tox

Pilot Study (10 chem)
In vivo results available
(use DEREK, TOPKAT, misc. SAR models)

80-90% prediction accuracy:
Propose use with single confirmatory assay

80% prediction accuracy:
Propose use with reduced confirmatory assay

Poor prediction accuracy (10%)
No reduction in animal use

Implementation would lead to 37% reduction in use of animals

Structure-based Screening & Prioritization:

Data

Apply existing SAR model

Data

Chemical(s) of concern

Data

Develop new SAR model

Data

Place in context of existing data and understanding

Data

Mine existing data for structural & biofunctional analogs
Chemistry-based Data Mining & Exploration:

- Chemical(s) of concern
- Structural analogs
- Property analogs
- Biological/mechanistic analogs
- Chemical-specific data

Structure-Activity Relationships
Reports: NTP is converting study reports into an electronic format which can be accessed from the website. These reports are made available as soon as they have been converted.

Data Searches: The NTP has been loading study information into databases and has developed applications to access this data from the web. There are two types of data mining searches:

- All types of data - search provides a way to find the various types of studies conducted on a test agent and has options to mine that data if it is available in electronic format.
- Bioassay pathology data mining search provides a way to access the pathology databases. It is also possible to search the historical control database and to view
NTP Database Search Home Page

Please note: This new NTP website is a "Work-in-Progress" project. Click here for a.

Search by CAS No. or all or part of the chemical name

Note: This search includes synonyms, but the search results will display the primary chemical name, the CAS number and the synonym name. For additional help, press the "Help" button in the top menu bar.

View a list of studies with available electronic data

Choose Study Type To Search Across Similar Studies

Note: This search capability is under construction. Currently only the pathology for the 2-year rodent studies stored in the Toxicology Data Management System (TDMS) since about 1983 is searchable. More than 200 studies are loaded into the database for searching and we continue to add to this set as time permits. The search looks for significant chemical-related toxicity data.

Enter a CAS number or Chemical Name:

CASRN or Chemical Name search

Off-site to search structure or analogs (NLM ChemID Plus)
### NTP Studies on 1,1,1,2-Tetrabromoethane

**Table Instructions and Notes:**
- Choose study type to view data for 1,1,1,2-Tetrabromoethane
- Not all agents have been studied in every study type. If there are no electronically available studies on 1,1,1,2-Tetrabromoethane, you may contact the NTP Central Files (cdm@niehs.nih.gov) to request available data from completed studies or the status for ongoing studies

#### Standard Toxicology & Carcinogenesis Studies

<table>
<thead>
<tr>
<th>Description of standard protocols</th>
<th>Reproductive Studies</th>
<th>Developmental Studies</th>
<th>Immunology Studies</th>
<th>Genetic Toxicity Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study C91016</td>
<td>No Reproductive Studies Available for this Chemical</td>
<td>No Developmental Studies Available for this Chemical</td>
<td>No Immunology Studies Available for this Chemical</td>
<td>In Vitro Study Data</td>
</tr>
<tr>
<td>Status: Rpt Complete</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Length: 3-Week</td>
<td></td>
<td></td>
<td></td>
<td>Description of standard protocols</td>
</tr>
<tr>
<td>Route: Gavage</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rats: FISCHER 344</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**View study results:** Genetic toxicity studies: Salmonella

**5 study categories**
Salmonella Study Overview

Current Search Criteria

<table>
<thead>
<tr>
<th>Chemical Name</th>
<th>CAS Number</th>
<th>Study Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,1,1,2-Tetrabromoethane</td>
<td>630-16-0</td>
<td>Salmonella</td>
</tr>
</tbody>
</table>

Table Instructions and Notes:
Click on the study number to view a summary of the results.
Standard NTP Protocol

<table>
<thead>
<tr>
<th>Study ID</th>
<th>Study Result</th>
<th>Year Completed</th>
</tr>
</thead>
<tbody>
<tr>
<td>A87711</td>
<td>Weak Positive</td>
<td>1993</td>
</tr>
</tbody>
</table>

Summary Call
View study details
**Summary Table**

<table>
<thead>
<tr>
<th>Study ID</th>
<th>Overall Result</th>
<th>Year Completed</th>
</tr>
</thead>
<tbody>
<tr>
<td>A87711</td>
<td>Weak Positive</td>
<td>1993</td>
</tr>
</tbody>
</table>

**Table Instructions and Notes:**
Click 'View Detailed Data' to proceed to the study data.

<table>
<thead>
<tr>
<th>Options</th>
<th>Strain</th>
<th>S9 Activation</th>
<th>S9 Species</th>
<th>Concentration</th>
</tr>
</thead>
<tbody>
<tr>
<td>View Detailed Data</td>
<td>TA100</td>
<td>Yes</td>
<td>Hamster</td>
<td>10% HLI</td>
</tr>
<tr>
<td></td>
<td>TA100</td>
<td>Yes</td>
<td>Rat</td>
<td>10% RLI</td>
</tr>
<tr>
<td></td>
<td>TA100</td>
<td>Yes</td>
<td>Hamster</td>
<td>30% HLI</td>
</tr>
<tr>
<td></td>
<td>TA100</td>
<td>Yes</td>
<td>Rat</td>
<td>30% RLI</td>
</tr>
<tr>
<td></td>
<td>TA100</td>
<td>No</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>TA1535</td>
<td>Yes</td>
<td>Hamster</td>
<td>10% HLI</td>
</tr>
<tr>
<td></td>
<td>TA1535</td>
<td>Yes</td>
<td>Rat</td>
<td>10% RLI</td>
</tr>
<tr>
<td></td>
<td>TA1535</td>
<td>Yes</td>
<td>Hamster</td>
<td>30% HLI</td>
</tr>
<tr>
<td></td>
<td>TA1535</td>
<td>Yes</td>
<td>Rat</td>
<td>30% RLI</td>
</tr>
</tbody>
</table>
### Strain: TA1538

<table>
<thead>
<tr>
<th>Dose (ug/Plate)</th>
<th>30% RLI (Negative)</th>
<th>30% HLI (Negative)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Std. Error</td>
</tr>
<tr>
<td>Vehicle Control</td>
<td>10</td>
<td>2.70</td>
</tr>
<tr>
<td>100</td>
<td>12</td>
<td>3.80</td>
</tr>
<tr>
<td>200</td>
<td>14</td>
<td>2.60</td>
</tr>
<tr>
<td>400</td>
<td>13</td>
<td>2</td>
</tr>
<tr>
<td>600</td>
<td>14</td>
<td>1.20</td>
</tr>
<tr>
<td>800</td>
<td>12</td>
<td>1.50</td>
</tr>
<tr>
<td>Positive Control</td>
<td>168</td>
<td></td>
</tr>
</tbody>
</table>

*Individual experiment results*

### Strain: TA97

<table>
<thead>
<tr>
<th>Dose (ug/Plate)</th>
<th>10% RLI (Negative)</th>
<th>30% RLI (Equivocal)</th>
<th>10% HLI (Negative)</th>
<th>30% HLI (Negative)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Std. Error</td>
<td>Mean</td>
<td>Std. Error</td>
</tr>
<tr>
<td>Vehicle Control</td>
<td>179</td>
<td>8.10</td>
<td>178</td>
<td>23.40</td>
</tr>
<tr>
<td>100</td>
<td>168</td>
<td>9.30</td>
<td>252</td>
<td>9.30</td>
</tr>
<tr>
<td>200</td>
<td>192</td>
<td>3.40</td>
<td>241</td>
<td>9.40</td>
</tr>
<tr>
<td>400</td>
<td>T</td>
<td>0</td>
<td>234</td>
<td>9.20</td>
</tr>
<tr>
<td>600</td>
<td>T</td>
<td>0</td>
<td>243</td>
<td>8.40</td>
</tr>
<tr>
<td>800</td>
<td>T</td>
<td>0</td>
<td>T</td>
<td>0</td>
</tr>
<tr>
<td>Positive Control</td>
<td>734</td>
<td>28</td>
<td>376</td>
<td>14.40</td>
</tr>
</tbody>
</table>
Individual experiment results

- Cannot download entire list of NTP chemicals and test summary data
- Cannot structure or substructure-search database
- Cannot download subsets of data:
  - list of TA98 pos data
  - list of all thyroid tumor carcinogens
- Cannot ask relational questions of data:
  - what chemicals are TA100 neg + TA98 pos?
  - list all chemicals with positive rat liver tumor findings in cancer bioassay that are also non-mutagenic

<table>
<thead>
<tr>
<th>Strain: TA1538</th>
<th>Dose (ug/Plate)</th>
<th>30% RL (Negative) Mean</th>
<th>Std. Error</th>
<th>30% HL (Negative) Mean</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vehicle Control</td>
<td>100</td>
<td>10.29</td>
<td>1.06</td>
<td>10.30</td>
<td>1.06</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>20.55</td>
<td>2.10</td>
<td>20.60</td>
<td>2.11</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>40.80</td>
<td>4.21</td>
<td>40.90</td>
<td>4.22</td>
</tr>
<tr>
<td></td>
<td>600</td>
<td>61.05</td>
<td>6.32</td>
<td>61.10</td>
<td>6.33</td>
</tr>
<tr>
<td></td>
<td>800</td>
<td>81.30</td>
<td>8.43</td>
<td>81.40</td>
<td>8.44</td>
</tr>
<tr>
<td>Positive Control</td>
<td>734</td>
<td>73.40</td>
<td>7.48</td>
<td>73.50</td>
<td>7.49</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Strain: TA97</th>
<th>Dose (ug/Plate)</th>
<th>10% RL (Negative) Mean</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vehicle Control</td>
<td>100</td>
<td>1.29</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>2.55</td>
<td>0.52</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>4.80</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>600</td>
<td>6.05</td>
<td>1.04</td>
</tr>
<tr>
<td></td>
<td>800</td>
<td>8.30</td>
<td>1.25</td>
</tr>
<tr>
<td>Positive Control</td>
<td>734</td>
<td>73.40</td>
<td>7.48</td>
</tr>
</tbody>
</table>
Archival vs. Relational
Short-Term Bioassays

NTP Database Site Map

Developmental Studies

Embryo/fetal malformations
“Shared standards are a huge flattener, because they both force and empower more people to communicate and innovate over much wider platforms.”

The World is Flat - Thomas Friedman, 2004.
Distributed Structure-Searchable Toxicity Public Database Network
Distributed
Structure-Searchable
Toxicity
Public
Database
Network

Chemical structure-annotation

Data standards and integration
DSSTox

The Distributed Structure-Searchable Toxicity (DSSTox) Database Network is a project of EPA's Computational Toxicology Program, helping to build a public data foundation for improved structure-activity and predictive toxicology capabilities. The DSSTox website provides a public forum for publishing downloadable, standardized toxicity data files that include chemical structures.

Recent Additions: 1Mar05

**New Database Additions:**
- FDA Maximum (Recommended) Daily Dose Database (FDAMDD) of 1217 pharmaceuticals - 1Mar05

**Expanded and modified versions:**
- Consolidated, updated Carcinogenic Potency Database - All Species (CPDBAS), 1451 compounds: 91 new records added to v2a
- CAS registry numbers added to EPAPFH and DBPCAN

**New Standard Fields added to all DSSTox files:**
- INChI (IUPAC/NIST Chemical Identifier) unique structure-text codes
- IUPAC systematic chemical names (generated by ACD/Name)
- Standard Toxicity Fields: StudyType, Species, Endpoint fields

**New Features of Site:**
- FTP Download Instructions for easy access to archived and new DSSTox data files
- New information pages: INChI, DSSTox Standard Toxicity Fields
- Links to External Public Databases adopting DSSTox standards: ISSCAN

DSSTox Databases:
- CPDBAS v2a 1451 1Mar05
- DBPCAN v2a 209 1Mar05
- EAPFH v2a 617 1Mar05
- FDAMDD_v1a 1217 1Mar05
- NCTERER_v2a 232 1Mar05

** new addition
### DSSTox Standard Chemical Fields:

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>STRUCTURE</td>
<td>Tested chemical, general form of chemical, active ingredient of formulation, representative isomer in mixture, representative component in mixture, representative component in mixture, monomer of polymer, simplified to parent</td>
</tr>
<tr>
<td>STRUCTURE_CASRN</td>
<td>Single chemical compound</td>
</tr>
<tr>
<td>STRUCTURE_CASRN Description</td>
<td>Macromolecule defined mixture or formulation</td>
</tr>
<tr>
<td>STRUCTURE_CASRN Description</td>
<td>Undefined mixture or formulation</td>
</tr>
<tr>
<td>STRUCTURE_CASRN Description</td>
<td>Unspecified or multiple forms</td>
</tr>
<tr>
<td>TestSubstance_CASRN</td>
<td>Defined organic inorganic organometallic</td>
</tr>
<tr>
<td>TestSubstance_Description</td>
<td>ChemicalNote</td>
</tr>
<tr>
<td>TestSubstance_Description</td>
<td>-Nature &amp; CAS of mixture components, -tautomers, -sterechemistry, -error in source info -replicate 2D, CAS, parent -quaternary ammonium</td>
</tr>
<tr>
<td>STRUCTURE_MolecularWeight</td>
<td>Defined organic inorganic organometallic</td>
</tr>
<tr>
<td>STRUCTURE_Formula</td>
<td>Parent, salt Na, Cl, etc complex HCl, H2O, mesylate, etc</td>
</tr>
<tr>
<td>STRUCTURE_ChemicalType</td>
<td></td>
</tr>
<tr>
<td>STRUCTURE_TestedForm_DEFINEDORGANIC</td>
<td></td>
</tr>
<tr>
<td>STRUCTURE_IUPAC</td>
<td></td>
</tr>
<tr>
<td>STRUCTURE_SMILES</td>
<td></td>
</tr>
<tr>
<td>STRUCTURE_Parent_SMILES</td>
<td></td>
</tr>
<tr>
<td>STRUCTURE_InChI</td>
<td></td>
</tr>
</tbody>
</table>
Integrating Diverse Databases from a Chemical Structure Perspective:

CPDB  |  DBPCAN  |  EPAFHM  |  NCTRER  ....

SAL CPDB  |  ChemClass DBP  |  ChemClass FHM  |  NCTRlogRBA
TD50 Rat  |  Concern Level  |  MOA  |  ER RBA
TD50Mouse |  Rationale  |  MOACONF  |  ChemClass ERB
Target Sites Rat Male  |  Analog  |  CLOGP  |  Activity Group ERB
Target Sites Rat Female  |  ChemName  |  LC50  |  Rationale
Target Sites Mouse Male  |  Analog  |  LC50NOTE  |  ChemClass ERB
.... Other Species  |  ChemName  |  LC50RATIO  |  MeanChem
AnalogCAS  |  MIXMOA  |  Class ERB RBA
AnalogSMILES  |  TOXINDEX  |  LogP
FATS  |  BEHAVIOR

F1, F2, ...F6
Integrating Diverse Databases from a Chemical Structure Perspective:

- CPDB
- DBPCAN
- EPAFHM
- NCTRER

**Standard Chemical Fields**

**Standard Tox Fields:** StudyType, Species, Endpoint, Route

**Activity Fields:** Dose, ActivityCategory, SummaryCall, ...

- TD50Mouse
- Target Sites Rat Male
- Target Sites Rat Female
- Target Sites Mouse Male
- Other Species
- Concern Level
- Rationale
- Rational Source
- Analog
- ChemName
- AnalogCAS
- AnalogSMILES
- MOACONF
- CLOGP
- LC50
- LC50NOTE
- LC50RATIO
- MIXMOA
- TOXINDEX
- FATS
- BEHAVIOR
- ChemClass ERB
- Activity Group ERB
- Rationale
- ChemClass ERB
- MeanChem
- Class ERB RBA
- LogP
- F1, F2, ... F6
DSSTox Database Design:

**Toxicology**
- Toxicologists
- Biologists
- Risk Assessors
- Domain Expertise

**Chemistry**
- SAR Modeling
- Machine Learning
- Comp. Chemistry
- 3D QSAR

**Toxicological Data**

**Chemical Structures & Properties**

**Context**

**Utility for SAR**

**Relevance**
NTPGTZ_v1a_1921
SAL_TA100, rat S9, hamster S9, w/o S9, sum
SAL_TA102, ...
SAL_TA104
SAL_TA1535
SAL_TA1537
SAL_TA97
SAL_TA98
vag = vagina;
vs = vascular system.

adr = adrenal gland;
on = bone;
ci = clitoral gland;
eso = esophagus;
ey = ear/Zymbal's gland;
gal = gall bladder;
hag = hardeman gland;
hmo = hematopoietic system;
ki = kidney;
lgi = large intestine;
liv = liver;
lun = lung;
meo = mesovarium;
mgl = mammary gland;
mix = mixture;
myc = myocardium;
nas = nasal cavity
nrv = nervous system;
or = oral cavity;
ova = ovary;
pan = pancreas;
per = peritoneal cavity;
pit = pituitary gland;
pre = preputial gland;
vsc = vascular system.

NTP_GTZ_v1a_1921
SAL_TA100, rat S9, hamster S9, w/o S9, sum
SAL_TA102, ...
SAL_TA104
SAL_TA1535
SAL_TA1537
SAL_TA97
SAL_TA98
vag = vagina;
vs = vascular system.
IMMTOX: Immunotoxicity Test Battery

- 88 chemicals
- 18 immunotox measures
- Summary calls
- Usage categories

AntibodyResponse
NaturalKillerCells
LymphocyteProliferation
MixedLeukocyteResponse
LeukocyteCount
ThymusBodyWt
SpleenBodyWt
Lipopolysaccharide
DelayedTypeHypersensitivity
CytotoxicTLymphocyte
SurfaceMarkers
ContactSensitization
HostResistanceAssays (6)
Up next ...

- **IMMTOX**: Immunotox Battery Database *Expanded from 1992 publication reporting a battery of immunotox results for 88 chemicals, most from the NTP.*

- **NCTRAR**: FDA’s National Center for Toxicological Research - Androgen Receptor Binding Database *Androgen receptor relative binding affinities tested in a common in vitro assay for 202 chemicals, provided with chemical class-based structure activity features.*

- **NTPGTZ**: NIEHS National Toxicology Program Gene-Tox Database (E. Zeiger) *Battery of genetic toxicity test results for over 1900 chemicals from historical NTP studies.*

- **UNLVSS**: UniLever’s Skin Sensitization Database *Skin sensitization results for over 200 chemicals from Unilever studies.*

- **Structure-Index Files:**
  - **IRISSI**: EPA’s Integrated Risk Information System (IRIS)
  - **EPAHPV**: EPA’s High Production Volume Chemicals
  - **NTPTSI**: National Toxicology Program’s Test Results
Total Records: 6625
Total Unique Records: 3967 (no replicates)
Chemical Class Overlaps:

Amino acids
Bases, Nucleoside
Benzenes
Carbocycle
Carbohydrates
Functional groups
Nitro
Nitrosamine
Organophosphor
Organohalides
Polyhalides
Heterocycles
Naphthalenes
Natural products

% frequency
Activity/Chemical Class Overlaps:
DSSTox Coordination/Collaborations:

- LeadScope Simulation Plus BioRad ACD/Labs
- Stanford Machine Learning Datasets
- NCI Govt. Databases
- PubChem
- ToxML
- Tox Data Standards
- NTP
- CEBS
- ArrayExpress, GEO, CTD, ...

Models
Data Files
Structure relational searching
Standard Chemical Fields
# Toxicity Experimental Data → Summary Data:

## Toxicity Content Model

<table>
<thead>
<tr>
<th>Test Substance</th>
<th>NamesID</th>
<th>SubstanceType</th>
<th>Descriptors</th>
<th>QA</th>
<th>Properties</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study Type</td>
<td>In Vivo</td>
<td>GenoTox</td>
<td>In Vitro</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test System</td>
<td>Animal Info</td>
<td>Feed Info</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test Procedure</td>
<td>Duration</td>
<td>Frequency</td>
<td>Route Exposure</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test Results</td>
<td>Effects Summary</td>
<td>Lesion Summary</td>
<td>Incident Rate</td>
<td>Survival Rate</td>
<td></td>
</tr>
</tbody>
</table>

- **Activity categories**
- **Potency categories**
- **Mode of action categories**
- **Summary calls**

---

ToxML

DSSTox

Standard Chemical &Tox Fields

- Toxicology
- Safety
- Environment
Toxicity Experimental Data $\rightarrow$ Summary Data:

**Toxicity Content Model**

- **ToxML**

- **Intermediate toxicity classifications for SAR**
  - Activity categories
  - Potency categories
  - Mode of action categories
  - Summary calls

**DSSTox Standard Chemical &Tox Fields**

```
<table>
<thead>
<tr>
<th>Test Substance</th>
<th>Names/ID</th>
<th>SubstanceType</th>
<th>Descriptors</th>
<th>QA</th>
<th>Properties</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study Type</td>
<td>In Vivo</td>
<td>GenoTox</td>
<td>In Vitro</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test System</td>
<td>Animal Info</td>
<td>Feed Info</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test Procedure</td>
<td>Duration</td>
<td>Frequency</td>
<td>Route Exposure</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test Results</td>
<td>Effects Summary</td>
<td>Lesion Summary</td>
<td>Incident Rate</td>
<td>Survival Rate</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gender: M/F</td>
<td>Gender: M/F</td>
<td>Gender: M/F</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>Negative</td>
<td>Clear evidence</td>
<td>Tissue Site</td>
<td>Dosage</td>
<td></td>
</tr>
<tr>
<td>Negative</td>
<td></td>
<td>Clear evidence</td>
<td>Lesion Type</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clear evidence</td>
<td></td>
<td>Some evidence</td>
<td>Lesion Site</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Some evidence</td>
<td></td>
<td>No evidence</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No evidence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Endpoint:
- TDLo
- TD50
- LOAEL
- MTD....
```
850,000 Substances from “Legacy” Public Sources
650,000 Unique Structures
10,000,000 BioAssay Results from NCI / DTP

structure/sub-structure searching
links to protein structures
links to PubMed
bioactivity screens
similar activity profiles
bioassay data
Googling for INChIs; A remarkable method of chemical searching

Peter Murray-Rust\textsuperscript{a}, Henry S. Rzepa\textsuperscript{b} and Yong Zhang\textsuperscript{a}

\textsuperscript{a}Unilever Centre for Molecular Informatics, University of Cambridge, UK, \textsuperscript{b}Department of Chemistry, Imperial College London, SW7 2AY,
EPA Data Challenges:

- “Electronification” of historical data
- Content annotation of unstructured data
- Chemical structure-annotation
- Data standardization & integration

CBI data

DSSTox

EPA-wide data

Government-wide data

- InChI text annotation
- EPA Structure-browser
- Collaborations with NCI, ATSDR, ECB, FDA, NIEHS

World-Wide Web
NIEHS/National Center for Toxicogenomics: Chemical Effects in Biological System Knowledge-Base (M. Waters and J. Fostel)

Gene expression
Gene Pathways
Gene Function
Proteomics
Metabonomics
Historical NTP Toxicity data
DSSTox / CEBS Collaboration:
Part 1 – DSSTox Annotation of CEB
DSSTox / CEBS Collaboration:
Part 1 – DSSTox Annotation of CEB

Structure-searchability
Analog searching
Cross domain searching

DSSTox Standard Chemical Fields

CEBS Vision - Bioinformatics to Knowledge

Associated Data
- Database of array proteomics and toxicology data on chemicals, drugs and stressors
- Database on genes and gene groups relevant to environmental disease
- Compendia of functional gene groups with associated pathways and networks
- Dictionaries and metadata

External Links
- NTP and other tox. databases
- Links
- Pathways
- Function
- Proteins
- Genomic resources
- Genes/protein descriptions

Query
- Compound/structure
- Effects
- Genes/protein functional groups
DSSTox / CEBS Collaboration:

Part 1 – DSSTox Annotation of CEB
Part 2 – Link CEBS to DSSTox Database Network

DSSTox Database Network +

Gene expression
Gene Pathways
Gene Function
Proteomics
Metabonomics

DSSTox Toxicity Data Files
DB1 DB2 DB3 DB4 DB5 DB6 DB7

DSSTox Standard Chemical Fields

CEBS Vision - Bioinformatics to Knowledge

DSSTox Toxicity Data Files

NTP and other Tox. Database
LITL
Pathways
Function

Gene/Protein Functional Groups
Genomic Resources
Genome/Protein Descriptions

Gene/Protein Dictionary/Structure
Effects

Database of SNP and Mutants Relevant to Environmental Disease
Compendia of Functional Gene Groups with Associated Pathways and Networks
Dictionary and Metadata

Database on Genes and Gene Groups Relevant to Environmental Disease

Retrieve
Store and Convert

Network +
Public Genomic Databases

Entrez is the integrated interface for NCBI for the mapping of Genes, Genomes, Chemicals, Protocols, and other...
Predictive Toxicology

SAR Predictions based solely on chemical structures & properties: 
- MCASE
- TOPKAT
- DEREK

Toxicity predictions based on gene expression profiles: 
- Gene-Logic
- Iconix
“Glocalization” of Predictive Tox Models

Chemical Structures

Mech 1
Mech 2
Mech 3
Mech 4
Mech 5

Toxicity Endpoint

Chemical reactivity

Biological attributes

chemistry structure
toxicity biology
Biological spectra analysis: Linking biological activity profiles to molecular structure

Anton F. Fliri*, William T. Loging, Peter F. Thadeio, and Robert A. Volkmann*

Pfizer Global Research and Development, Groton, CT 06340

Communicated by Larry E. Overman, University of California, Irvine, CA, October 25, 2004 (received for review September 4, 2004)

Establishing quantitative relationships between molecular structure and broad biological effects has been a longstanding challenge in science. Currently, no method exists for forecasting broad biological activity profiles of medicinal agents even within narrow boundaries of structurally similar molecules. Starting from the premise that biological activity results from the capacity of small organic molecules to modulate the activity of the proteome, we set out to investigate whether descriptor sets could be developed for measuring and quantifying this molecular property. Using a 1,576-compound database, we show that percent inhibition values determined at single high drug concentration in a battery of in vitro assays representing a cross section of the proteome, provide precise molecular property descriptors that identify the structure of molecules. When broad biological activity of molecules is represented in spectra form, organic molecules can be sorted by quantifying differences between biological spectra. Unlike traditional structure–activity relationship methods, sorting of molecules using biospectra comparisons does not require knowledge of a molecule’s putative drug targets. To illustrate this finding, we selected as starting point the biological activity spectra of clotrimazole and tioconazole because their putative target, lanosterol demethylase (CYP51), was not included in the bioassay array. Spectra similarity obtained through profile similarity measurements and hierarchical clustering provided an unbiased means for establishing quantitative relationships between chemical structures and biological activity spectra. This methodology, which we have termed biological spectra analysis, provides the capability not only of sorting molecules on the basis of biospectra similarity but also of predicting simultaneous interactions of new molecules with multiple proteins.

Approximate “proteomic” signature

1576 compounds
X 92 assays

Azole cluster by biospectra similarity
Cerep: In vitro bioactivity profiles

PASS (Prediction of Activity Spectra for Substances)

- Expanded chemical "properties" in relation to biological activity
- Databases & models need to extend beyond pharmaceuticals to environmental/industrial chemicals spanning toxicity space
NIH/NCGC Roadmap: nihroadmap.nih.gov

Small Molecules High-Throughput Screening Initiative

- > 200K “small molecules”
- 7 dose dilutions
- > 200 bioactivity & cell-based assays

Sample “toxicity” chemical space:
- NTP chemicals
- EPA pesticides, inerts
- EPA HPV Chemicals
- DSSTox
- NCI/ChemNavigator

Reference dataset of toxicity-related chemicals with structures & bioactivity profiles
Toxico-Chemoinformatics: Data Standardization, Integration, Exploration

“In Silico” Predictions

Biochemical “target” assays

In Vitro assays

In Vivo whole animal studies
Toxico-Chemoinformatics: Data Standardization, Integration, Exploration

Calculated structures, properties

Biochemical “target” assays

Receptors, enzymes, proteins

In Vitro assays

Short-term tests

Cell-based assays

Genomics

Tissue, organs

Whole animal studies

Chronic, acute