Next Generation Computational Chemistry Tools to Predict Toxicity of CWAs

William (Bill) Welsh
welshwj@umdnj.edu

Prospective Funding by DTRA/ J STO-CBD
A State-wide, Regional and National Resource

< www.ebCTC.org >

Funded with support from the U.S. EPA
Major Research Thrusts

- MENTOR-DORIAN Computational Toxicology System that spans the Source->Dose->Outcome continuum
- The Environmental Bioinformatics Knowledge Base (ebKB: www.ebCTC.org)
- *ArrayTrack*: toxicological bioinformatics platform to process genomics, proteomics and metabonomics data
- Hepatocyte Metabolic Model for Xenobiotics
- *ChemTox*, a suite of chem-informatics tools for toxicant identification & characterization
MENTOR & DORIAN
Address the Source-to-Outcome Continuum

MENTOR:
Modeling ENvironment for TOtal Risk

DORIAN:
Dose-Response Information Analysis

Adapted from chart by R. Calderon, USEPA/NHEERL, 2003
The environmental bioinformatics Knowledge Base (ebKB) serves as a comprehensive compendium of tools, databases, and literature.
ArrayTrack Suite of Bioinformatics Tools

Data input and QC

Data Import

Normalization methods

Normalization

Gene Selection

Apply to

Data Exploration

Expression pattern bar charts

Apply to

Interpretation

Gene Ontology

Pathway analysis

Identify significant genes

• Volcano Plot (considering both $p$ and fold-change)
• P-value Plot (considering false positives/negatives)

Individual gene analysis
ChemTox, an Integrated Suite of Cheminformatics Tools

Predictive Molecular Toxicology

Protein Modeling

Chemical Modeling

Molecule-Surface Interactions < skin, water, polymer >

Database Mining/Pattern Recognition

Jan 2007 CBIS Conference
Computational Screening Paradigm

- Priority Setting -

untested chemicals → Rejection Filters
Structural Alerts → Hits

Quantitative Prediction (QSARs) → Active

Inactive → Classification Models
Hierarchical Screening Framework

- addresses the need to minimize *false negatives* and *uncertainties*
- recognizes that no single computational model is adequate

**Tier I**
- Rejection Filters
  - MW
  - Structural alerts

**Tier II**
- Classification Models
  (active/inactive assignment)
  - Pharmacophores
  - Decision Forest
  - Shape Signatures

**Tier III**
- Quantitative Predictions
  - PLS & ANN QSAR models
  - PNN QSAR model
  - Receptor Docking

**Tier IV**
- Knowledge-Base Approach
  - Other priority setting factors
  - Human expert knowledge
**Decision Forest**

- Improved classification by combining independent Decision Tree models -

**Key Features**

- Combining several independent yet predictive trees reduces misclassification
- DF structure permits assessment of prediction confidence
- Each tree consists of simple 'If-Then' branches, hence the DF is extremely fast

![Decision Forest Diagram](image-url)
**Schematic of Hierarchical Framework**

- addresses the need to minimize *false negatives* and *uncertainties* -

<table>
<thead>
<tr>
<th>Tier I</th>
<th>Rejection Filters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tier II</td>
<td>Classification Models (active/inactive assignment)</td>
</tr>
<tr>
<td>Tier III</td>
<td>Quantitative Predictions</td>
</tr>
<tr>
<td>Tier IV</td>
<td>Knowledge-Base Approach</td>
</tr>
</tbody>
</table>

**Tier I**
- MW
- Structural alerts

**Tier II**
- Pharmacophores
- Decision Forest
- Shape Signatures

**Tier III**
- PLS & ANN QSAR models
- PNN QSAR model
  - Receptor Docking

**Tier IV**
- Other priority setting factors
- Modifying Tiers I, II, III
- Human expert knowledge
Shape Signatures Tool

START

Small molecule or Protein binding pocket

PROCESSING

Ray tracing to generate the raw data

OUTPUT

1D and 2D Shape Signatures

Jan 2007

CBIS Conference
*Shape Signatures Tool*

molecules are compared by subtracting their histograms

Small $\text{Diff}$ value means that two molecules have similar shape and polarity
**Shape Signatures Software Tool & Chemical Databases**

**Searchable Shape Signatures Databases**

- 3+ million commercially available organic compounds
- 40,000 Natural Products
- Hazardous Chemicals (pesticides, nerve agents, mustards, psychotropic agents, other real or potential CWAs, TI Cs)
- PDB-extracted ligands
Chemical → Target Protein → Mechanisms

Protein Data Bank (PDB): World Repository of ~35,000 Protein-Ligand Crystal Structures (http://www.rcsb.org/pdb/)

In this page you can select organisms. The protein ID and the 2D images of the ligands will be displayed in a table form.

Species/Protein Family

Shape Signatures of PDB-extracted ligands

Here are the results obtained by searching for HUMAN:

Protein Structure

Species/Protein Family

Jan 2007

CBIS Conference 16
<table>
<thead>
<tr>
<th>Molecules</th>
<th>Target Protein</th>
<th>Mechanism</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query Molecule</td>
<td>Matching PDB Ligands</td>
<td>Protein’s Binding Site</td>
</tr>
<tr>
<td>Shape Sigs PDB Ligands</td>
<td>Target Proteins</td>
<td>Links to Toxicity Pathways</td>
</tr>
<tr>
<td>Public Databases</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Identifying Problem Chemicals

**Simulant Chemicals**

- "Red flag" chemicals

**Query Chemical**

- In-house databases
- CWAs (real/potential)
- Pharmaceuticals & their biproducts
- Industrial chemicals

**Query Compound**

**Shape Signatures**

<table>
<thead>
<tr>
<th>Chemical Structure</th>
<th>CAS Number</th>
<th>1D score</th>
<th>2D score</th>
<th>Toxicity Data (mg/kg-day)</th>
<th>Endpoint(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Shape Signatures

- Key Features -

- Fast
  screens large databases in secs

- Extensible
  works with any kind or number of molecular species

- Portable
  works on any platform

- Versatile
  broad utility, multiple databases
Schematic of Hierarchical Framework

- addresses the need to minimize *false negatives* and *uncertainties* -

**Tier I**
- Rejection Filters
  - MW
  - Structural alerts

**Tier II**
- Classification Models (active/inactive assignment)
  - Pharmacophores
  - Decision Forest
  - Shape Signatures

**Tier III**
- Quantitative Predictions
  - PLS & ANN QSAR models
  - PNN QSAR model
  - Receptor Docking

**Tier IV**
- Knowledge-Base Approach
  - Other priority setting factors
  - Modifying Tiers I, II, III
  - Human expert knowledge
**Building QSAR Models**

**Types of Molecular Descriptors**

<table>
<thead>
<tr>
<th>Type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constitutional</td>
<td>Molecular composition ($M_w$, # of atoms/bonds, # of H-bond donors/acceptors)</td>
</tr>
<tr>
<td>Topological</td>
<td>2-D structural formula (Kier-Hall indices, extent of branching)</td>
</tr>
<tr>
<td>Geometrical</td>
<td>3-D structure of molecule (molecular volume, solvent accessible surface area, polar and non-polar surface area)</td>
</tr>
<tr>
<td>Electrostatic</td>
<td>Charge distribution (atomic partial charges, electronegativities)</td>
</tr>
<tr>
<td>Quantum Mechanical</td>
<td>Electronic structure (HOMO-LUMO energies, band gap, dipole moment)</td>
</tr>
</tbody>
</table>
Desirable Features of Methods and Models

- predictions should be fast
- produces linear or non-linear models (i.e., relationship between observed toxicities and calculated molecular features may be non-linear)
- models should be physically meaningful, interpretable, and assume parametric form

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>PLS/ MVR</td>
<td>**</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>ANN</td>
<td>*</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>PNN</td>
<td>**</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Polynomial Neural Network (PNN)

- combines best features of linear multivariate models (parametric form) and ANN models (nonlinearity) -

Polynomial Neural Network

- Produces linear or non-linear QSAR models in parametric form
- User control of model complexity
- Insensitive to irrelevant variables and outliers
- Yields predictive models, even for sparse or noisy data sets
- Trains rapidly, thus amenable to large data sets
- Automatically selects best models
- Customizable to fit user's needs
Polynomial Neural Network (PNN)

1) PNN generates parametric solutions of any desired order ‘n’:

\[
\text{Act.} = w_1(SA) + w_2(V) + w_3(\mu) + \ldots
\]

\[
\text{Act.} = w_1(SA) + w_2(V)^2 + w_3(\mu)^3 + \ldots
\]

\[
\text{Act.} = w_1(SA)^2 + w_2(V) + w_3(\mu)^2 + \ldots
\]

\[
\text{Act.} = w_1(SA)^0 + w_2(V) + w_3(\mu)^2 + \ldots
\]

\[
\text{Act.} = w_1(SA) + w_2(V)^2 + w_3(\mu)^2 + \ldots
\]

2) PNN selects best solutions:

\[
\text{Act.} = w_1(SA) + w_2(V)^2 + w_3(\mu)^3 + \ldots
\]

\[
\text{Act.} = w_1(SA) + w_2(V)^2 + w_3(\mu)^2 + \ldots
\]
Thank You!

welshwj@umdnj.edu