Cheminformatics and Chemical Information

Matt Sundling

Advisor: Prof. Curt Breneman
Department of Chemistry and Chemical Biology/Center for Biotechnology and Interdisciplinary Studies
Rensselaer Polytechnic Institute
Rensselaer Exploratory Center for Cheminformatics Research

http://reccr.chem.rpi.edu/

Many thanks to Theresa Hepburn!
Upcoming lecture plug:

Prof. Curt Breneman
DSES Department
November 15th

Advances in Cheminformatics: *Applications in Biotechnology, Drug Design and Bioseparations*
**Cheminformatics** is about collecting, storing, and analyzing [usually large amounts of] chemical data.

- Pharmaceutical research
- Materials design
- Computational/Automated techniques for analysis
- Virtual high-throughput screening (VHTS)

---

**QSAR** - quantitative *structure-activity* relationship

![Diagram showing the relationship between molecular structures, model, and activity]
AAACCTCATAGGAAGCA
TACCAGGAATTACATCA
...

Molecular Structures ↔ Model ↔ Activity
**Activity**: bioactivity, ADME/Tox evaluation, hERG channel effects, p-456 isozyme inhibition, anti-malarial efficacy, etc…

<table>
<thead>
<tr>
<th>Molecule</th>
<th>D1</th>
<th>D2</th>
<th>…</th>
<th>Activity (IC50)</th>
</tr>
</thead>
<tbody>
<tr>
<td>molecule #1</td>
<td>21</td>
<td>0.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>molecule #2</td>
<td>33</td>
<td>2.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>molecule #3</td>
<td>10</td>
<td>0.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Molecular Structures → Descriptors → Model → Activity
Goal: Minimize the error: \( \sum(y_i - f(x_i)) \)

Regression Models: linear, multi-linear, higher order functions, neural networks, etc…

Molecular Structures → Descriptors → Model → Activity
DESCRIPTORS:

- Structural Descriptors
- Physiochemical Descriptors
- Topological Descriptors
- Geometrical Descriptors
- Constitutional Descriptors
- Electrostatic Descriptors
- Quantum-chemical Descriptors
- Thermodynamic Descriptors

Molecular Structures $\xrightarrow{\text{Describing}}$ Model $\xleftarrow{\text{Activity}}$
QSAR (of birds) --> Quantitative Bird Structure-Property Relationship (QBSPR):

Bird Species --> Descriptors --> Model --> Activity
QSAR (of birds):

QBSPR: I want to understand the relationship between species and flight performance...

QSAR: I want to understand the relationship between compound and Acetylcholinesterase (AChE) inhibition...

Bird Species → Descriptors ← Model ← Activity
QSAR (of birds):

QBSPR: What is flight performance? Performance = muscle efficiency * flight time as (P = Efficiency*Time)

QSAR: What is AChE inhibition? IC50 is a measure of the concentration required for 50% inhibition.
QSAR (of birds):

QBSPR: What features or characteristics are related to what we want to predict? What differences between bird species affect what we are trying to determine (flight performance)?

QSAR: What features or characteristics of a molecule, affects it’s ability to inhibit AChE?
DESCRIPTORS (of birds):
DESCRIPTORS (of birds):

Bird Descriptors:
- Height, weight, ‘size’
- Color, Shape of beak, length of talons
- Bone structure
- Muscle structure
- Biokinetics, energetics, wake structure

MODEL:
\[ P \propto \text{feather length & weight} \]
DESCRIPTORS (of birds):

MODEL:

\[ P \propto \text{feather length} \& \text{weight} \]

BETTER MODEL:

\[ P \propto \frac{\text{(wing span)}^2}{\text{wing surface area}} \& \frac{\text{body mass}}{\text{wing surface area}} \]

*(wing surface area) is a latent descriptor. **Latent** = dormant, potential or hidden

QUESTIONS: Does your model make sense? What does it say about performance (P)? How does ‘conformation’ affect your model?
QSAR (of molecules):

What features of a molecule are related to my activity? What descriptors can capture that information?
DESCRIPTORS:

Structural Descriptors
Physiochemical Descriptors
Topological Descriptors
Geometrical Descriptors

Constitutional Descriptors
Electrostatic Descriptors
Quantum-chemical Descriptors
Thermodynamic Descriptors

• Analysis requires appropriate data
• Analysis requires relevant data
• Hierarchy of descriptors (data content)

• Same descriptors may not work in all situations (e.g. molecular weight predicts freezing point depression)

\[ \Delta T_f = K_f \times m \]
**DESCRIPTOR HIERARCHY:**

- Wave function of system or PE hypersurface
- ‘3D descriptors’ (e.g. stereo-chemical descriptors)
- ‘2D descriptors’ (e.g. connectivity matrices)
- Molecular formula

**Hierarchy of descriptors (data content)**

- Molecular Structures
- Descriptors
- Model
- Activity
**Molecular Surface Properties**

- **Electronic Properties**
  - Electrostatic Potential
  
  \[ EP(r) = \sum_{\alpha} \frac{Z_{\alpha}}{|r - R_{\alpha}|} - \int \frac{\rho(r')dr'}{|r - r'|} \]

  \[ K(r) = -(\psi* \nabla^2 \psi + \psi \nabla^2 \psi^*) \]

  \[ G(r) = -\nabla \psi^* \cdot \nabla \psi \]

  \[ \nabla \rho \cdot \mathbf{N} \]

  \[ L(r) = -\nabla^2 \rho(r) = K(r) - G(r) \]

  \[ PIP(r) = \sum \frac{\rho_i(r) |\epsilon_i|}{\rho(r)} \]

  \[ F^+(r) = \rho \text{HOMO}(r) \]

  
  - Electronic Kinetic Energy Density
  - Electron Density Gradients
  - Laplacian of the Electron Density
  - Local Average Ionization Potential
  - Bare Nuclear Potential (BNP)
  - Fukui function
Why use Electron Density-Derived Molecular Descriptors?

• Motivations
  – Electron Density Distributions represent molecular properties that are key to biological activities

• Enabling Technologies
  – Fast methods (TAE/RECON) for obtaining electron density-derived properties

• Encoding schemes
  – Surface Property distributions (Histograms, Wavelets, Dixels)
  – Shape/Property hybrid distributions (PEST)

• Synergies
  – Complementary to topological descriptors
Surface Property Distribution Histograms (RECON/TAE) Descriptors

Molecular surface property distributions can be represented as RECON/TAE histogram bin descriptors.
Wavelet Decomposition:
- Creates a set of coefficients that represent a waveform.
- Small coefficients may be omitted to compress data.

Wavelet Reconstruction:
- 16 coefficients of the WCD vector represent surface property densities with >95% accuracy.
original signal

DWT

TAE histogram descriptors

reconstructed signal

iDWT

wavelet coefficient descriptors (WCDs)
PEST: Molecular Shape/Property Hybrid Encoding

- **PEST** (Property-Encoded Surface Translation)
  - Adds shape information to encode the spatial relationships of surface properties
PEST Molecular Ray Tracing Algorithm

QuickTime™ and a decompressor are needed to see this picture.
Understanding and Interpretation

Structural Descriptors
Physiochemical Descriptors
Topological Descriptors
Geometrical Descriptors

wavelet coefficient descriptors (WCDs)
Conclusions: *These things are important!*

1. Descriptors classes
2. Selection of descriptors
3. Selection of descriptor *representation*