Electron density derived descriptors in ADME/Tox screening

Presented by
N. SUKUMAR
Curt M. Breneman, Mark J. Embrechts, Kristin P. Bennett and Dechuan Zhuang

http://www.drugmining.com/

Copyright, 2005 © Rensselaer Polytechnic Institute.
ADMET Property Prediction: Challenges in Medicinal Chemistry

Multiple-parameter optimization of lead structures

- Other parameters: patent position, chemical synthesis
- The greatest hurdle: ADMET properties.
Traditional Drug Discovery Scheme

- Potency
- Absorption
- Distribution
- Metabolism
- Excretion
- Toxicity

Lead → Drug

- Potency
- Absorption
- Distribution
- Metabolism
- Excretion
- Toxicity
In silico prediction of ADME properties
Electron Density Derived descriptors

- TAE (Transferable Atom Equivalent)
- WCD (Wavelet Coefficient Descriptors)
- PEST (Property-Encoded Surface Translator) Hybrid Shape-Property Descriptors
- Autocorrelation of Molecular Surface Properties
Electron Density Derived Molecular Surface Properties

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

- Electronic Kinetic Energy Density
  
  \[ K(r) = - (\psi^* \nabla^2 \psi + \psi \nabla^2 \psi^*) \]
  
  \[ G(r) = - \nabla \psi^* \cdot \nabla \psi \]

- Electron Density Gradients
  
  \[ \nabla \rho \cdot \mathbf{N} \]

- Laplacian of the Electron Density
  
  \[ L(r) = - \nabla^2 \rho(r) = K(r) - G(r) \]

- Local Average Ionization Potential
  
  \[ PIP(r) = \sum \frac{\rho_i(r)}{\rho(r)} |\varepsilon_i| \]

- Bare Nuclear Potential (BNP)

- Fukui function
  
  \[ F^+(r) = \rho_{\text{HOMO}}(r) \]
Reconstruction Method

- RECON algorithm for rapid reconstruction of molecular charge densities and molecular electronic properties
- based on Bader’s quantum theory of Atoms In Molecules
- A library of atomic charge density fragments -- corresponding to structurally distinct atom types -- has been built in a form that allows for the rapid retrieval of the fragments and molecular assembly
- Associated with each atomic charge density fragment in the library is a data file which contains
  - information describing topological features of the atomic charge density used to orient the fragments into their proper molecular space orientations
  - atomic charge density-based descriptors encoding electronic and structural information relevant to the chemistry of intermolecular interactions.
Theory of Atoms in Molecules

Definition of an Atom in a Molecule:

◆ An atom is the union of an attractor and its basin
◆ Each atom contains one \((\text{and only one})\) nucleus, which is the attractor of its electron density distribution \(\rho(r)\)
◆ Every atom is bounded by an atomic surface of zero flux
◆ Atoms defined in this way satisfy the virial theorem
◆ They have properties that are approximately additive and transferable from one molecule to another.
Molecular input
- PDB
- Tripos MOL2
- SIMS DRA
- SMILES
- SDF

Determine atom types and assign closest match from atom type library according to the priority:
- Perfect match
- Ring size differs
- Hybridization of nearest neighbor does not match
- Atomic number of nearest neighbor does not match
- Hybridization of atom does not match
- For monovalent atom, hybridization of nearest neighbor differs
- Atomic number of atom does not match

Combine densities of atomic fragments
Compute predicted molecular properties

http://www.drugmining.com/
# RECON Timing Performance

<table>
<thead>
<tr>
<th>Hardware</th>
<th>SGI 300 MHz Octane</th>
<th>1.7 GHz Intel Pentium</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>O/S</strong></td>
<td>IRIX64</td>
<td>Windows 2000</td>
</tr>
<tr>
<td><strong>Test dataset</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td># of molecules</td>
<td></td>
<td></td>
</tr>
<tr>
<td># of molecules</td>
<td>196</td>
<td>69</td>
</tr>
<tr>
<td>File Format</td>
<td>mixed</td>
<td>mixed</td>
</tr>
<tr>
<td>User CPU (sec.)</td>
<td>23.5</td>
<td></td>
</tr>
<tr>
<td>System CPU (sec.)</td>
<td>1.4</td>
<td></td>
</tr>
<tr>
<td>Total CPU (sec.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Misc. small molecules</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td># of molecules</td>
<td>1650</td>
<td>1641</td>
</tr>
<tr>
<td>File Format</td>
<td>SDF</td>
<td>SMILES</td>
</tr>
<tr>
<td>User CPU (sec.)</td>
<td>102.7</td>
<td>122.3</td>
</tr>
<tr>
<td>System CPU (sec.)</td>
<td>44.5</td>
<td>45.9</td>
</tr>
<tr>
<td>Total CPU (sec.)</td>
<td>650.7</td>
<td>757.2</td>
</tr>
<tr>
<td><strong>Proteins</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td># of molecules</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td>File Format</td>
<td>PDB</td>
<td>PDB</td>
</tr>
<tr>
<td>User CPU (sec.)</td>
<td>186.8</td>
<td>194.5</td>
</tr>
<tr>
<td>System CPU (sec.)</td>
<td>194.5</td>
<td>194.5</td>
</tr>
<tr>
<td><strong>NCI</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td># of molecules</td>
<td>42,689</td>
<td>42,689</td>
</tr>
<tr>
<td>File Format</td>
<td>SDF</td>
<td>SDF</td>
</tr>
<tr>
<td>User CPU (sec.)</td>
<td>2327.2</td>
<td>2327.2</td>
</tr>
<tr>
<td>System CPU (sec.)</td>
<td>1131.0</td>
<td>1131.0</td>
</tr>
</tbody>
</table>
PEST Shape/Property Hybrid descriptors

- A TAE property-encoded surface is subjected to internal ray reflection analysis.

- A ray is initialized with a random location and direction within the molecular surface and reflected throughout inside the electron density isosurface until the molecular surface is adequately sampled.

- Molecular shape information is obtained by recording the ray-path information, including segment lengths, reflection angles and property values at each point of incidence.


PEST Shape/Property Hybrid descriptors

- Segment length and point-of-incidence value form 2D-histogram
- Each bin of 2D-histogram becomes a hybrid descriptor

Diazapam

EP Surface

EP vs Length 2D histogram

C. Matthew L. Sundling (COMP 345) PEST vs. CoMFA: A comparative study of two 3D-QSAR technologies (4:00 PM, Thursday, March 17 -- Room 7B)
General Modeling protocol

DATASET

Training set
  Bootstrap sample k
    Training
    Validation
      Learning Model
      Tuning / Prediction
  Test set
    Predictive Model
      Prediction
Traditional Regression approaches

- Minimize the training errors $\sum (y_i - f(x_i))$

Overfitting problem (HTS data with noise):
- “perfect” prediction for training set
- Poor prediction for unknown data
Support Vector Regression

Minimize the Generalization error:

\[ \min_{w, b, \xi_i, \xi_i^*} \left( C \sum_{i=1}^{l} (\xi_i + \xi_i^*) + \frac{1}{2} \|w\|^2 \right) \]

- Parameter \( C \) controls tradeoff between error and capacity
- Minimizing \( \|w\| \) controls capacity of linear function

\( \varepsilon \)-insensitive loss function:

\[ L_\varepsilon(y - f(x)) := \min(0, |y - f(x)| - \varepsilon) \]

Avoid overfitting by controlling the model complexity
Variable Selection using SVM

The basic idea is very simple:

– Construct a series of sparse linear SVM that exhibit good generalization.

– Construct the subset of variables having nonzero weights in the linear models.

– Then use this subset of variables in nonlinear SVM to produce the final regression or classification function.

The method exploits the fact that linear SVM with 1-norm regularization inherently performs variable selection as a side-effect of minimizing capacity in the SVM model.
Nonlinear Support Vector Machine

- Construct linear function in higher-dimensional space.
- Mapping done using Kernel functions.
- Typically use RBF Kernel
- Regression function is

\[ f(u) = \sum_{i=1}^{\ell} \left( (\alpha_i - \alpha_i^*) K(u, x_i) \right) - b \]
Human Intestinal Absorption

Drugs Diffusion Mechanisms

- For most drugs, the preferred route of administration is by oral ingestion
- Oral drugs are absorbed across a series of separate barriers
- Diffusion mechanisms:
  - Passive diffusion
  - Active diffusion
Different barriers

A series of separate barriers (epithelial layer is the most dominant barrier)

Drugs

Mucus Gel Layer

Intestinal Epithelial Cells

Lamina Propria

Endothelium of Capillaries

Be absorbed
What is Caco-2 cell line?

- Derived from a human colorectal carcinoma
- With remarkable morphological and biochemical similarity to the small intestinal columnar epithelium
- The apparent permeability coefficient (Papp) has been shown to correlate well with %HIA
- Extremely useful for mechanistic studies of drug absorption and as an absorption screening assay for pre-clinical drug selection.
- The absorption of a drug compound through the human intestinal cell line is an important property for potential drug candidates.
- Caco-2 cell line is a widely used model for drug absorption.
- Accurately measuring this property, however, can be very costly and time-consuming.
- The use of QSPRs is an attractive alternative to experimental measurement.
**Caco-2 Bagged SVM Model**
(RBF Kernel, 14 features selected by SVM/LP)

- Human intestinal cell line
- Predicts drug absorption
- 27 molecules with tested permeability
- 718 descriptors generated
  - Electronic TAE
  - Shape/Property (PEST)
  - MOE

Train $R^2 = 0.93$
Test $R^2 = 0.83$

![Graph showing Observed vs Predicted log Papp (cm/sec)]
Caco-2 – 14 Features

- **Hydrophobicity** - a.don
- **Size and Shape** - ABSDRN6, SMR.VSA2, ANGLEB45

  Large is bad. Flat is bad. Globular is good.
- **Polarity** - PEOE.VSA....: negative partial charge good.

Each star represents a descriptor. Each ray is a separate bootstrap. The area of a star represents the relative importance of that descriptor. Descriptors shaded cyan have a negative effect. Unshaded ones have a positive effect.
Modeling Fathead Minnow Toxicity

- 375 Diverse Organic Compounds
- Acute Toxicity in terms of $-\log LD_{50}$ (negative logarithm of mM concentration causing 50% fatality in fathead minnow population within 96 hours of exposure)
- 147 RECON 2D descriptors
- Randomly take out 125 molecules as the external test set
- Build Support Vector Regression models with all descriptors and with 20 selected descriptors
- Check model significance using Y-scrambling
  - Scramble the response values of the training set: the correlation coefficient measures the “messiness”
  - Build a model from the scrambled training data
  - Predict the external test set using the model
  - Repeat 50 times
- Compare the performance of the real model and the scrambled models
Model with 20 Selected Descriptors

Model with 147 Descriptors

Models with Y-Scrambling

Modeling Results

- $r^2 = 0.8064$
- $R^2 = 0.8011$
- $RMSE = 2.79$
- $q^2 = 0.5465$
- $Q^2 = 0.5365$

- $r^2 = 0.8677$
- $R^2 = 0.8695$
- $RMSE = 2.31$
- $q^2 = 0.5453$
- $Q^2 = 0.5329$
Starplots of 20 Selected Descriptors

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIEPM</td>
<td>13.02%</td>
</tr>
<tr>
<td>Lap13</td>
<td>-9.70%</td>
</tr>
<tr>
<td>Del(Rho)NM</td>
<td>-7.22%</td>
</tr>
<tr>
<td>Energy</td>
<td>-6.30%</td>
</tr>
<tr>
<td>SIKIA</td>
<td>6.01%</td>
</tr>
<tr>
<td>Fuk6</td>
<td>5.45%</td>
</tr>
<tr>
<td>SIGA1</td>
<td>5.29%</td>
</tr>
<tr>
<td>SIEPA1</td>
<td>-4.75%</td>
</tr>
<tr>
<td>SIEPA7</td>
<td>-4.73%</td>
</tr>
<tr>
<td>Del(G)NA9</td>
<td>4.47%</td>
</tr>
<tr>
<td>PIP2</td>
<td>3.94%</td>
</tr>
<tr>
<td>Del(Rho)NMMax</td>
<td>3.57%</td>
</tr>
<tr>
<td>Lap15</td>
<td>3.34%</td>
</tr>
<tr>
<td>SIEPA7</td>
<td>3.21%</td>
</tr>
<tr>
<td>Del(G)NA9</td>
<td>2.94%</td>
</tr>
<tr>
<td>PIP10</td>
<td>2.94%</td>
</tr>
<tr>
<td>Del(K)NA5</td>
<td>-2.71%</td>
</tr>
<tr>
<td>EP7</td>
<td>2.69%</td>
</tr>
<tr>
<td>EP9</td>
<td>2.62%</td>
</tr>
<tr>
<td>PIP3</td>
<td>1.91%</td>
</tr>
<tr>
<td>Lap10r</td>
<td>1.91%</td>
</tr>
</tbody>
</table>

- Solubility, surface area important
- Del(G)N, Del(Rho)N describe polarizability/hydrophobicity
- EP correlated with solvation behavior
Autocorrelation of Molecular Surface Properties

Warfarin

\[ p(x) \text{ property at point } x \]
\[ N \text{ # of products in distance interval} \]
\[ d \text{ distance interval} \]

RECON Autocorrelation Descriptors

- Derived from RECON binned histogram descriptors
- Based on Gasteiger’s Autocorrelation Formula
  \[ A(R_{xy}) = \frac{1}{n} \sum_{x,y} P_y P_x \]
- Binned by distance between atoms \( x \) and \( y \) (RAD) or by topological distance between these atoms (TRAD)
- Reflects 3-dimensional information content but can be calculated almost as rapidly as 2-dimensional descriptors
- Calculated for each type of standard RECON descriptor
- Higher bin number => greater distance between atom

RECON Autocorrelation of G: Assorted Molecules

RECON Topological Autocorrelation of G: Assorted Molecules
### Modeling Fathead Minnow Toxicity

Kernel PLS with Subjective Feature Selection

<table>
<thead>
<tr>
<th>Descriptors</th>
<th>r² (training)</th>
<th>q² (test)</th>
<th># Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>RECON</td>
<td>0.821</td>
<td>0.611</td>
<td>100</td>
</tr>
<tr>
<td>Moe2d, RECON</td>
<td>0.851</td>
<td>0.61</td>
<td>100</td>
</tr>
<tr>
<td>RECON, RAD</td>
<td>0.836</td>
<td>0.613</td>
<td>100</td>
</tr>
<tr>
<td>RECON, TRAD</td>
<td>0.839</td>
<td>0.632</td>
<td>100</td>
</tr>
<tr>
<td>RAD, TRAD</td>
<td>0.766</td>
<td>0.541</td>
<td>100</td>
</tr>
<tr>
<td>Moe2d, Moe3d, RECON</td>
<td>0.859</td>
<td>0.625</td>
<td>100</td>
</tr>
<tr>
<td>Moe2d, RECON, TRAD</td>
<td>0.864</td>
<td>0.63</td>
<td>100</td>
</tr>
<tr>
<td>RECON, RAD, TRAD</td>
<td>0.857</td>
<td>0.587</td>
<td>100</td>
</tr>
<tr>
<td>All 5 sets</td>
<td>0.875</td>
<td>0.622</td>
<td>100</td>
</tr>
</tbody>
</table>

**Training Set:** 250 molecules  
**Test Set:** 125 molecules
MOE 2-D & 3-D descriptors

Objective MOE 2d, MOE 3d Training Set
$r^2 = 0.83137$

Subjective RECON, TRAD Training Set
$r^2 = 0.83886$

Objective MOE 2d, MOE 3d Test Set
$q^2 = 0.36987$

Subjective RECON, TRAD Test Set
$q^2 = 0.36764$
Histograms or wavelet encoding of surface properties gives TAE property descriptors

- Surface histograms can represent property distributions with 70-75% accuracy when 10-20 histogram bins are used

- >95% accuracy with 16 wavelet coefficients

The $\rho(r) = 0.002 \text{ e/au}^3$ isosurface corresponds roughly to the molecular van der Waals surface.
Wavelet Decomposition

Photographic images are routinely compressed using Discrete Cosine Transforms: Joint Photographic Experts Group (JPEG)

<table>
<thead>
<tr>
<th>Original image</th>
<th>High Resolution JPEG</th>
<th>Medium Resolution JPEG</th>
<th>Low Resolution JPEG</th>
</tr>
</thead>
<tbody>
<tr>
<td>514 KB</td>
<td>26 KB</td>
<td>11 KB</td>
<td>6.6 KB</td>
</tr>
</tbody>
</table>

In this example, nearly 90-fold data compression is achieved by retaining only the highest amplitude coefficients.

We use Discrete Wavelet Transforms (DWT) to encode quantum chemical information content.
Electron Density-Derived Molecular Properties: Wavelet Coefficient Descriptors (WCD)

Wavelet Decomposition:
- Creates a set of coefficients that represent a waveform.
- Small coefficients may be omitted to compress data.

Wavelet Surface Property Density Reconstruction:
16 coefficients from S7 and D7 portions of the WCD vector represent surface property densities with >95% accuracy.

Increased property accuracy of descriptors provides more predictive models.

1024 raw wavelet coefficients capture PIP distribution on molecular surface.
Blood-Brain Barrier Modeling

**WCD vs Traditional Descriptors**

16 TAE Wavelet Coefficient Descriptors  
16 Traditional 2D & 3D Descriptors

Both cases used GAPLS feature selection and three latent variables. Cross-validation was performed using bootstrapping techniques. All data values shown were predicted as unknowns using 80/20 training/test groupings.

Classification of HERG blockers

Approximate location of mutation(s).
The number of mutations are shown in white.
ROC curve (median)

Sensitivity vs. Specificity

Confusion matrix

100 molecules in validation set

409 molecules in training set

False positives

False negatives

Negatives

Positives

Confusion matrix

<= 3

51

4

> 3

20

25
Acknowledgements

• Breneman Research Group:
  – Dechuan Zhuang
  – Matt Sundling
  – Bill Katt
  – Minghu Song
  – Lingling Shen
  – Bo Jiang
  – Larry Lockwood

• Embrechts Research Group

• Bennett Research Group

See: www.drugmining.com for more details