Application of Chemometric and QSAR Approaches to Scoring Ligand Receptor Binding Affinity

Alexander Tropsha
Laboratory for Molecular Modeling
University of North Carolina at Chapel Hill

Collaborators: C. Breneman, W. Deng, N. Sukumar (RPI); A. Golbraikh, J. Feng, Y.D.Xiao (UNC)
OUTLINE

• Representation and scoring of ligand-receptor interactions in the context of structure-based design

• Chemometric and QSAR Approaches to Scoring
  – Complementarity between ligands and active sites in multidimensional descriptor space
  – QSAR modeling of diverse ligands of different receptors
  – Database mining
Scoring function is the crucial component of structure-based drug design.
### Molecular Representations and Scoring Functions for Ligand-Receptor Interaction

<table>
<thead>
<tr>
<th>Representation</th>
<th>Scoring function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Molecular mechanics</td>
<td>Force-Field based</td>
</tr>
<tr>
<td>Molecular mechanics</td>
<td>Empirical estimates of intermolecular interaction terms (VALIDATE, SCORE)</td>
</tr>
<tr>
<td>Chemical atom types</td>
<td>Statistical, or knowledge-based (PMF, SMoG, BLEEP)</td>
</tr>
<tr>
<td>Multiple Chemical Descriptors</td>
<td>Chemical similarity, QSAR (QSBR)</td>
</tr>
</tbody>
</table>
Representation of Ligand-Receptor Interface in Chemometric Space

• Both ligands and active sites are represented independently by multidimensional chemical descriptors
• Chemical similarity measures are used to establish complementarity between ligands and their receptors
• QSAR approaches are used to establish correlations between descriptors of ligands and their binding affinity to diverse receptors
Datasets


Serine Proteases: 16
Metalloproteases: 15
L-arabinose binding protein: 9
Endothiapepsin: 11
Other proteins: 17


Aspartic Proteases: 18 complexes
Serine Proteases: 20
Metalloproteases: 22
Human Carbonic Anhydrase II (HCA, 19 Complexes) Sugar-Binding Proteins (14 Complexes)
Endothiapepsin (11 Complexes)
Purine Nucleoside Phosphatase (PNP, 5 Complexes)
Other Proteins (10 Complexes)
SCORING...BUT NO DOCKING: QSAR ANALYSIS OF LIGANDS TO DIFFERENT RECEPTORS: MOLECULAR DIVERSITY
Transferable Atom Equivalents
(TAE Descriptors)*

◆ A library of atomic charge density fragments has been built in a form that allows for the rapid retrieval of the fragments and molecular assembly.

◆ Atoms in the library correspond to *structurally distinct atom types*

◆ Associated with each atomic charge density fragment in the TAE library are a set of descriptors.

◆ The atomic descriptors in the TAE library are calculated *ab initio* at a high level of theory (HF 6-31+G*)

Mapping of PMF dataset from high-dimensional to 2D TAE Descriptor Space using SVD/TNPACK
(Xie, Tropsha, Schlick. JCICS, 1999, 2000; 167-177)
Algorithm to establish complementarity between ligands and their receptors in multidimensional TAE descriptor space using variable selection

Randomly Select a Subset of $M_{var}$ Descriptors

Calculate pairwise distances in selected subspace based on ligand TAE

Calculate pairwise distances in selected subspace based on protein TAE

Build correlations between corresponding pairwise distances

Calculate $r^2$ in selected descriptor subspace
RR vs. LL Distances in the Entire TAE Space (2016 points)

y = 1.0771x
R^2 = 0.5028
LL vs RR Distances in Selected Descriptors Space of 20 TAE descriptors (2016 points)

Ligand-Protein Distance Pairs After Variable Selection (Nselected=10)

$y = 1.2244x$

$R^2 = 0.7059$
Most frequently occurring TAE descriptors in 35 best models
Chemical meaning of selected descriptors

- Electrostatic interactions, induced-dipole interactions and hydrogen bonding
- Hydrophobic surface area
QSAR ANALYSIS OF LIGANDS OF DIFFERENT RECEPTORS

OBJECTIVE:
BUILD QSAR MODELS USING INFORMATION ABOUT LIGAND STRUCTURES ONLY
NORMALIZATION:  

\[ X_{ij}^n = \frac{X_{ij} - X_{j,\text{min}}}{X_{j,\text{max}} - X_{j,\text{min}}} \]

- \( X_{ij} \) and \( X_{ij}^n \) are the non-normalized and normalized j-th descriptor values for compound \( i \), \( X_{j,\text{min}} \) and \( X_{j,\text{max}} \) are the minimum and maximum values for j-th descriptor.

MODEL VALIDATION:  
Prediction of binding energies for the test set compounds. Comparison of predicted and observed binding energies.
2D Descriptors

Set of descriptors includes topological features of compounds

Molecular connectivity indices
Molecular shape indices
Wiener number
Platt number
Information indices
Graph radius and diameter
etc.

Example: Molecular connectivity indices

\[ 0 \chi = \sum_{i=1}^{N} (a_i)^{-0.5} \]
\[ 1 \chi = \sum_{\text{all edges}} (a_{i_1} a_{i_2})^{-0.5} \]
\[ n^{-1} \chi = \sum_{\text{all (n-1)-edge subgraphs}} (a_{i_1} a_{i_2} \ldots a_{i_v})^{-0.5} \]

The number of descriptors exceeds the number of compounds

Randomly select a subset of descriptors

LEAVE-ONE-OUT CROSS-VALIDATION

Exclude a compound

Predict activity $y$ of the excluded compound as the weighted average of activities of 1 to $K$ nearest neighbors

$$\hat{y} = \frac{\sum_{i} y_i \exp(-d_i)}{\sum_{i} \exp(-d_i)}$$

Calculate the predictive ability ($q^2$) of the “model”

$$q^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (\bar{y} - y_i)^2}$$

Modify descriptor subset

Select the best QSAR model for $nvar$ and $K$

Zheng, W. and Tropsha, A. 
*JCICS.*, 2000; 40; 185-194
PRACTICING SAFE QSAR

Regression

\[ y' = a' y + b' \]

\[ a' = \frac{\sum (y_i - \bar{y})(\bar{y}_i - \bar{y})}{\sum (y_i - \bar{y})^2} \]

\[ b' = \bar{y} - a' \bar{y} \]

Correlation coefficient

\[ R = \frac{\sum (y_i - \bar{y})(\bar{y}_i - \bar{y})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (\bar{y}_i - \bar{y})^2}} \]

Regression through the origin

\[ \bar{y}' = k'y \]

\[ k' = \frac{\sum y_i \bar{y}_i}{\sum y_i^2} \]

Coefficients of determination

\[ R^2 = 1 - \frac{\sum (\bar{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} \]

\[ R^2_0 = 1 - \frac{\sum (y_i - \bar{y}_i')^2}{\sum (y_i - \bar{y})^2} \]

CRITERIA

\[ q^2 > 0.5; R^2 > 0.6; \]

\[ k \text{ or } k' \approx 1.0; R^2_0 \text{ or } R^2_0' \approx R^2 \]
QSAR ANALYSIS OF PMF DATASET: RESULTS

59 COMPOUNDS

TRAINING SET: 50
TEST SET: 9

$q^2=0.737$
$R^2_{pred}=0.850$
$R_0^2=0.847$
$k=1.137$
$S.E.=0.479$
$F=39.58$

TRAINING SET: 46
TEST SET: 13

$q^2=0.584$
$R^2_{pred}=0.757$
$R_0^2=0.743$
$k=1.108$
$S.E.=0.509$
$F=34.22$

TRAINING SET: 39
TEST SET: 20

$q^2=0.510$
$R^2_{pred}=0.718$
$R_0^2=0.716$
$k=0.850$
$S.E.=0.653$
$F=45.74$
QSAR ANALYSIS OF LIGANDS OF DIFFERENT RECEPTORS: OBSERVED VS. PREDICTED BINDING ENERGIES

TRAINING SET: 50
TEST SET: 9

\[ y = 1.0183x \]
\[ R^2 = 0.7393 \]
\[ y = 1.0198x + 0.0637 \]
\[ R^2 = 0.7393 \]
\[ y = 0.8551x \]
\[ R^2 = 0.841 \]
\[ y = 0.7835x - 2.914 \]
\[ R^2 = 0.8497 \]

TRAINING SET: 46
TEST SET: 13

\[ y = 1.0391x \]
\[ R^2 = 0.5978 \]
\[ y = 0.8586x - 7.3812 \]
\[ R^2 = 0.6287 \]
\[ y = 1.1078x \]
\[ R^2 = 0.7436 \]
\[ y = 0.9911x - 3.8447 \]
\[ R^2 = 0.7568 \]
QSAR ANALYSIS OF LIGANDS TO DIFFERENT RECEPTORS

111 RECEPTOR-LIGAND COMPLEXES

111 LIGANDS

ELIMINATION OF IDENTICAL COMPOUNDS

95 COMPOUNDS
QSAR ANALYSIS OF SMoG2001 dataset: BEWARE OF $q^2$!
### QSAR Analysis of Ligands to Different Receptors: Results

#### 95 Compounds

<table>
<thead>
<tr>
<th>N</th>
<th>Training set</th>
<th>Test set</th>
<th>$q^2$</th>
<th>$R^2$</th>
<th>$R_0^2$</th>
<th>k</th>
<th>$s^2$</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>88</td>
<td>7</td>
<td>0.67</td>
<td>0.96</td>
<td>0.96</td>
<td>0.95</td>
<td>0.50</td>
<td>122.8</td>
</tr>
<tr>
<td>2</td>
<td>82</td>
<td>13</td>
<td>0.63</td>
<td>0.84</td>
<td>0.80</td>
<td>0.95</td>
<td>1.52</td>
<td>57.5</td>
</tr>
<tr>
<td>3</td>
<td>74</td>
<td>21</td>
<td>0.60</td>
<td>0.82</td>
<td>0.78</td>
<td>0.97</td>
<td>1.98</td>
<td>88.3</td>
</tr>
<tr>
<td>4</td>
<td>67</td>
<td>28</td>
<td>0.67</td>
<td>0.70</td>
<td>0.68</td>
<td>0.95</td>
<td>2.83</td>
<td>61.4</td>
</tr>
<tr>
<td>5</td>
<td>60</td>
<td>35</td>
<td>0.57</td>
<td>0.67</td>
<td>0.66</td>
<td>0.97</td>
<td>2.93</td>
<td>72.1</td>
</tr>
<tr>
<td>6</td>
<td>52</td>
<td>43</td>
<td>0.50</td>
<td>0.61</td>
<td>0.61</td>
<td>0.96</td>
<td>1.75</td>
<td>64.3</td>
</tr>
<tr>
<td>7</td>
<td>44</td>
<td>51</td>
<td>0.54</td>
<td>0.57</td>
<td>0.57</td>
<td>0.97</td>
<td>2.04</td>
<td>70.7</td>
</tr>
</tbody>
</table>
QSAR ANALYSIS OF LIGANDS TO DIFFERENT RECEPTORS: OBSERVED VS. PREDICTED BINDING ENERGIES

TRAINING SET: 88
TEST SET: 7

Predicted log(Kd) vs. Observed log(Kd)

- y = 0.9733x - 0.1802
  - R² = 0.673
- y = 0.9958x
  - R₀² = 0.6726

TRAINING SET: 67
TEST SET: 28

Predicted log(Kd) vs. Observed log(Kd)

- y = 1.2418x + 1.9386
  - R² = 0.6938
- y = 0.9979x
  - R₀² = 0.6662

- y = 1.1454x + 1.5642
  - R² = 0.7024
- y = 0.9467x
  - R₀² = 0.6794

- y = 1.0051x + 0.3972
  - R² = 0.9609
- y = 0.9528x
  - R₀² = 0.9577

- y = 0.9958x
  - R₀² = 0.6726
QSAR ANALYSIS OF LIGANDS TO DIFFERENT RECEPTORS: OBSERVED VS. PREDICTED BINDING ENERGIES

TRAINING SET: 44
TEST SET: 51

\[ y = 1.123x + 0.8693 \]
\[ R^2 = 0.5464 \]

\[ y = 1.0115x \]
\[ R_0^2 = 0.5408 \]

\[ y = 0.8957x - 0.5852 \]
\[ R^2 = 0.5731 \]

\[ y = 0.9691x \]
\[ R_0^2 = 0.5689 \]
Activity randomization to assess model robustness

None of the random models was acceptable
CONCLUSIONS

• Chemometric and QSAR approaches to representation and scoring of protein ligand interaction provide an important supplement to existing structure based methods

• Based on the description of active site atoms one can search for complementary ligands in chemical databases

• Ligand affinity can be predicted from multi-target QSAR models of ligands in the absence of the receptor structure

• (Optimal) binding free energy is defined by ligand structure alone?
ACKNOWLEDGMENTS

• UNC associates
  Former:
  – Weifan ZHENG
  – Sung Jin CHO
  – Xin Chen
  – Stephen CAMMER

Protein folding group:
  – John GRIER

  Jun FENG
  – Shuxing ZHANG
  – David BOSTICK
  – Bala KRISHNAMOORTHY
  – Ruchir SHAH
  – Sagar KHARE

QSAR group:
  – Alex GOLBRAIKH
  – Yun-De XIAO
  – Min SHEN
  – Scott OLOFF
  – Yuanyuan QIAO

• Collaborators:
  – C. Breneman (RPI)
  – D. Bonchev (Texas A&M)
  – R. Hormann (R&H)
  – C. Reynolds R&H→Ortho)
  – V. Gombar (GSK)
  – E. Gilford (Pfizer)

Funding
  – NIH
  – Millenium
  – Rohm & Haas
  – Ortho McNail
  – GSK
  – Pfizer