Lecture 18

- Quantitative Structure Property and Activity Relations
  - Multiple Linear Regression (MLR) and Partial Least Squares (PLS)
  - K-Nearest Neighbors (KNN)
  - Support Vector Machines (SVM)
- Virtual Screening
- Pharmacophore Mapping
Relate Chemical Structure and Biological Activity

Chemical Structure

Biological Activity
Transformation-Invariant, Problem-Optimized Numerical Description

Chemical Structure

Biological Activity
Machine Learning Calculates Activity from Numerical Description
Multiple Linear Regression (MLR) and Partial Least Squares (PLS)

- Descriptors of chemical structure is generally a vector of \( n \) variables collected for \( m \) small molecules:
  \[
  X = \begin{bmatrix}
  x_{11} & \cdots & x_{1n} \\
  \vdots & \ddots & \vdots \\
  x_{m1} & \cdots & x_{mn}
  \end{bmatrix}
  \]

- Biological activity or chemical property is generally a scalar collected for \( m \) small molecules:
  \[
  Y = \begin{bmatrix}
  y_1 \\
  \vdots \\
  y_m
  \end{bmatrix}
  \]

- Multiple Linear Regression (MLR) finds a vector of \( n \) weights \( W = \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix} \) so that \( \|X \times W - Y\| \) is minimized.
- Compute pseudo inverse of \( X \) using Singular Value Decomposition (SVD).
- Compute \( W = X^{-1} \times Y \).
- PLS is an extension of MLR where both sets \( X \) and \( Y \) are projected into a hyperspace. PLS regression is particularly suited when the matrix of predictors has more variables \( n \) than observations \( m \), and when there is linear dependence between some variables. Standard MLR will fail in these cases.
K-Nearest Neighbor (KNN) Models

- In pattern recognition, the KNN is a method for classifying objects based on closest training examples in the feature space.
- The KNN algorithm is amongst the simplest of all machine learning algorithms: an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors.
- \( k \) is a positive integer, typically small. If \( k = 1 \), then the object is simply assigned to the class of its nearest neighbor.
- The compound of interest is described by \( X_\star = [x_{\star 1}, \ldots, x_{\star n}] \).
- Find the \( k \) molecules \( X_{KNN} = \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{k1} & \cdots & x_{kn} \end{bmatrix} \) from your training data with smallest \( \|X_\star - X_i\| \).
- For these \( k \) molecules compute the average \( y_\star = \frac{1}{k} \sum_{i=1}^{k} y_i \) as prediction of activity/property.
Machine Learning Techniques for Pattern Recognition

<table>
<thead>
<tr>
<th></th>
<th>Computer</th>
<th>Human Brain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computational units</td>
<td>1 CPU, $10^5$ gates</td>
<td>$10^{11}$ neurons</td>
</tr>
<tr>
<td>Storage units</td>
<td>$10^9$ bits RAM, $10^{10}$</td>
<td>$10^{11}$ neurons, $10^{14}$ synapses</td>
</tr>
<tr>
<td>Cycle time</td>
<td>$10^{-6}$ sec</td>
<td>$10^{-3}$ sec</td>
</tr>
<tr>
<td>Bandwidth</td>
<td>$10^9$ bits / sec</td>
<td>$10^{14}$ bits / sec</td>
</tr>
<tr>
<td>Neuron updates / sec</td>
<td>$10^5$</td>
<td>$10^{14}$</td>
</tr>
</tbody>
</table>

Original <-> Pattern <-> Prediction

6 December 2011 © Jens Meiler
Supervised Machine Learning

- Extracting knowledge from given data samples in order to \textit{generalize} a classification model
- Derive \textit{patterns} out of massive data sets while comparing results with known data

![Supervised Machine Learning Diagram]
Supervised Machine Learning

- Extracting knowledge from given data samples in order to generalize a classification model
- Derive patterns out of massive data sets while comparing results with known data
Structure of Artificial Neurons

- Biological Neuron
- Artificial Neuron

![Diagram of a biological neuron and an artificial neuron with weighted inputs and an activation function.](image-url)
Structure of Artificial Neural Networks

- Biological Neural Networks
  - Human nervous system has $10^{11}$ neurons, each connected to as many as $10^5$ other neurons, $10^{14}$ synapses

- Artificial Neural Network
  - Input Layer
  - Hidden Layer
  - Output Layer
  - Outputs
Data Processing in an Artificial Neural Network

- n-dimensional input vector is projected into m-dimensional space
- in this process typical pattern within the input information are recognized
- $S^{j+1} = \text{sigmoid}(\sum w_{ij} \cdot S^j_i)$
Mapping Descriptor Space into Hyperspace
Support Vector Machine

- hyperplane with a maximal margin in descriptor space
- defined by Support Vectors on edge of margin
- constructed in descriptor space by a Kernel function
Data Processing in a Support Vector Machine

- Test vector $x$
- Support vectors $x_1 \ldots x_n$
- Mapped vectors in hyperspace $\Phi(x_i), \Phi(x)$
- Dot product $\langle \Phi(x_i), \Phi(x) \rangle = k(x, x_i)$
- Weights $w_1 \ldots w_n$
- Decision function $\sigma(\Sigma w_i k(x, x_i))$
Transformation-Invariant, Problem-Optimized Numerical Description

Problem-Optimized Descriptor Set

I(s)

a) b) c) d) e) f)

Biological Activity
Receiver Operating Characteristic (ROC) and Enrichment

Predicted Activity | Biological Activity | True Positive Rate | False Positive Rate
--- | --- | --- | ---
1.00 | 1 | 0.2 | 0
0.95 | 1 | 0.4 | 0
0.83 | 0 | 0.4 | 0.2
0.71 | 1 | 0.6 | 0.2
0.69 | 1 | 0.8 | 0.2
0.54 | 0 | 0.8 | 0.4
0.33 | 0 | 0.8 | 0.6
0.27 | 1 | 1 | 0.6
0.10 | 0 | 1 | 0.8
0.03 | 0 | 1 | 1

**Graph:**
- Random Predictor
- Trained QSAR Model
- Optimal Predictor
## Receiver Operating Characteristic (ROC) and Enrichment

<table>
<thead>
<tr>
<th>Predicted Activity</th>
<th>Biological Activity</th>
<th>True Positive Rate</th>
<th>False Positive Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>1</td>
<td>0.2</td>
<td>0</td>
</tr>
<tr>
<td>0.95</td>
<td>1</td>
<td>0.4</td>
<td>0</td>
</tr>
<tr>
<td>0.83</td>
<td>0</td>
<td>0.4</td>
<td>0.2</td>
</tr>
<tr>
<td>0.71</td>
<td>1</td>
<td>0.6</td>
<td>0.2</td>
</tr>
<tr>
<td>0.69</td>
<td>1</td>
<td>0.8</td>
<td>0.2</td>
</tr>
<tr>
<td>0.54</td>
<td>0</td>
<td>0.8</td>
<td>0.4</td>
</tr>
<tr>
<td>0.33</td>
<td>0</td>
<td>0.8</td>
<td>0.6</td>
</tr>
<tr>
<td>0.27</td>
<td>1</td>
<td>1</td>
<td>0.6</td>
</tr>
<tr>
<td>0.10</td>
<td>0</td>
<td>1</td>
<td>0.8</td>
</tr>
<tr>
<td>0.03</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

### Graphs

- **Random Predictor**
- **Trained QSAR Model**
- **Optimal Predictor**
Receiver Operating Characteristic
ROC, Enrichment, and Slope

External DB
- total: 10
- active: 5

1:1 = ratio active : inactive

Filtered DB
- total: 4
- active: 3

3:1 = ratio active : inactive

Enrichment = \(\frac{TP}{TP + FP} \div \frac{P}{P + N} = \frac{TP \times N + TP \times P}{FP \times P + TP \times P} = \frac{3}{4/2} = 1.5\)

Slope = \(\frac{TP}{P} / \frac{FP}{N} = \frac{TP \times N}{FP \times P} = 3\)
Allosteric Modulators of mGluR$_5$ Bind to the 7TM Domain of the GPCR

Glutamate binding site

Site 1: MPEP / DFB / CDPPB / VU-29

Site 2: CPPHA binding site

6 December 2011 © Jens Meiler
ROSETTA constructed mGluR5 Comparative Model for 7TM domain
Based on crystal structure of β-2-Adrenergic Receptor (2RH1)
Sequence identity in the TM spans is ~12%, similarity is ~35%
From this model S806, S808, T810, E578 are hypothesized as potential interaction sites for CPPHA and will be tested experimentally
Treatment Strategies for CNS Disorders through Modulation of mGluR5

- Allosteric positive modulation (activation) of mGluR5 may ameliorate the symptoms of schizophrenia.

- Allosteric negative modulation of mGluR5 offers a potential treatment strategy of fragile X syndrome symptoms, a CNS disorder associated with autism spectrum disorders (ASD).
High-Throughput Screen yields 1387 PAMs and 345 NAMs of mGluR$_5$

- 150,000 compounds were tested for allostERIC modulation of mGluR$_5$ measuring receptor-induced intracellular release of calcium. 1,387 (0.94%) compounds were verified as PAMs of mGluR$_5$. 345 (0.23%) compounds were verified as NAMs of mGluR$_5$.

Optimizing the set of chemical descriptors for the given target

Number descriptors

- Molecular Weight
- Number H bond donors
- Number H bond acceptors
- XlogP
- Polar surface area
- Mean molecular polarizability
- Molecular dipole moment
- Aqueous solubility

vHTS Training Optimization (ROC curves)
Optimizing the set of chemical descriptors for the given target

Number descriptors

8

1252

vHTS Training Optimization (ROC curves)

True Positives (%)

False Positives (%)
Optimizing the set of chemical descriptors for the given target

Number descriptors

vHTS Training Optimization (ROC curves)
Optimizing the set of chemical descriptors for the given target

Number descriptors

vHTS Training Optimization (ROC curves)
Virtual Screen for Highly Active Compounds and Novel Leads

- A) True positive
- B) False negative
- C) False positive
- D) True negative

- Enrichment of Active Compounds by 43x
Experimental Results mGluR$_5$
Positive Allosteric Modulators

~450,000
ChemBridge

824 Compounds predicted with $EC_{50} < 1\mu M$ by QSAR model

232 Compounds (28.1%) were confirmed as mGluR$_5$ PAMs
Enrichment = $28.1\% / 0.96\% = 30$

Non-trivial scaffold modifications with mGluR5 PAM activity

6 December 2011
© Jens Meiler
Experimental Results mGluR$_5$ Negative Allosteric Modulators

ChemBridge

~750,000

749 Compounds with novel Scaffolds predicted with EC$_{50} < 10\mu$M by QSAR model

12 Compounds (3.6%) were confirmed as mGlur$_5$ NAMs
Enrichment = 3.6% / 0.23% = 16

VU0240790-4
EC$_{50} = 75$ nM

VU0360620-1
EC$_{50} = 124$ nM

6 December 2011 © Jens Meiler
Comparative Molecular Field Analysis (CoMFA) 1. Effect of Shape on Binding of Steroids to Carrier Proteins

Richard D. Cramer, III,* David E. Patterson, and Jeffrey D. Bunce

Contribution from Tripos Associates, 1699 South Hanley Road, St. Louis, Missouri 63144. Received January 5, 1988

Abstract: Comparative molecular field analysis (CoMFA) is a promising new approach to structure/activity correlation. Its characteristic features are (1) representation of ligand molecules by their steric and electrostatic fields, sampled at the intersections of a three-dimensional lattice, (2) a new “field fit” technique, allowing optimal mutual alignment within a series, by minimizing the RMS field differences between molecules, (3) data analysis by partial least squares (PLS), using cross-validation to maximize the likelihood that the results have predictive validity, and (4) graphic representation of results, as contoured three-dimensional coefficient plots. CoMFA is exemplified by analyses of the affinities of 21 varied steroids to corticosteroid- and testosterone-binding globulins. Also described are the sensitivities of results to the nature of the field and the definition of the lattice and, for comparison, analyses of the same data using various combinations of other parameters. From these results, a set of ten steroid-binding affinity values unknown to us during the CoMFA analysis were well predicted.
3D-QSAR CoMFA study of benzoxazepine derivatives as mGluR5 PAMs

3D-QSAR CoMFA study of benzoxazepine derivatives as mGluR₅ positive allosteric modulators

Edward W. Lowe Jr. a, Alysia Ferrebee b, Alice L. Rodriguez c, d, P. Jeffrey Conn c, d, Jens Meiler a, *

a Center for Structural Biology, Vanderbilt University, Nashville, TN 37232 6600, USA
b Department of Biology, Howard University, Washington, DC 20059, USA
c Department of Pharmacology, Vanderbilt University Medical Center, Nashville, TN 37232, USA
d Vanderbilt Program in Drug Discovery, Nashville, TN 37232, USA

ARTICLE INFO

Article history:
Received 30 April 2010
Revised 13 July 2010
Accepted 15 July 2010
Available online 30 July 2010

Keywords:
Metabotropic
mGluR₅
Allosteric
QSAR

ABSTRACT

Positive allosteric modulation of the metabotropic glutamate receptor subtype 5 was studied by conducting a comparative molecular field analysis on 118 benzoxazepine derivatives. The model with the best predictive ability retained significant cross-validated correlation coefficients of $q^2 = 0.58$ ($r^2 = 0.81$) yielding a standard error of 0.20 in pEC₅₀ for this class of compounds. The subsequent contour maps highlight the structural features pertinent to the bioactivity values of benzoxazepines.

© 2010 Elsevier Ltd. All rights reserved.
118 Benzoxazepine Derivatives with known PAM Activity were Fitted

Figure 3. (a) Represents the predictive versus experimental pEC$_{50}$ values for the CoMFA model, after geometry optimization, without region focusing. (b) Represents the predictive versus experimental pEC$_{50}$ values for the best CoMFA model.

Figure 4. Electrostatic (a) and steric (b) contour plots of the best CoMFA model, respectively. Areas favoring steric bulk are indicated by green polyhedra (contribution level 80%), while areas where steric bulk is unfavorable are represented by yellow polyhedra (contribution level 20%). Areas favoring electropositive values are indicated by blue polyhedra (contribution level 80%) while areas where electronegative values are favored are represented by red polyhedra (contribution level 20%). Contours were generated based on the results of the final, non-cross-validated PLS analysis.
BCL::PHARMMAP – Computes Partial Derivatives of Property vs. Structure

- The Algorithm

1. Input Scaffold
2. Monte Carlo Structure Generator
3. Derivatives $\Delta prop_i @ site_j$
4. QSAR Model $\Delta activity$

$sensitivity(prop_i @ site_j) \approx \left( \frac{\partial activity}{\partial prop_i @ site_j}_{prop_k \neq i @ site_i \neq j} \right) \frac{\Delta activity}{\Delta prop_i @ site_j}$
### BCL::PHARMMap – Generation of Chemical Derivatives

The diagram shows a scaffold molecule reacting with substituents to generate derivatives. The table below lists various substituents and their corresponding derivatives.

<table>
<thead>
<tr>
<th>Substituents</th>
<th>Derivatives</th>
</tr>
</thead>
<tbody>
<tr>
<td>-OH</td>
<td>-NH(CH₃)</td>
</tr>
<tr>
<td>-C(CH₃)₃</td>
<td>-CH₂-</td>
</tr>
<tr>
<td>-NH₂</td>
<td>-NH-</td>
</tr>
<tr>
<td>-H</td>
<td>-CH=CH₂</td>
</tr>
<tr>
<td>-COOH</td>
<td>-NH(CH₃)</td>
</tr>
<tr>
<td>-CH₃</td>
<td>-CH₂-</td>
</tr>
<tr>
<td>-F</td>
<td>-N</td>
</tr>
<tr>
<td>-CH=CH₂</td>
<td>-N-</td>
</tr>
<tr>
<td>-SH</td>
<td>-OCH₃</td>
</tr>
<tr>
<td>-OCH₃</td>
<td>-CH₂-</td>
</tr>
<tr>
<td>-CH₂CH₃</td>
<td>-CO-</td>
</tr>
<tr>
<td>-Br</td>
<td>-S-</td>
</tr>
<tr>
<td>-SC₃</td>
<td>-I</td>
</tr>
<tr>
<td></td>
<td>=CH-</td>
</tr>
<tr>
<td></td>
<td>=N-</td>
</tr>
</tbody>
</table>
Steric Bulk Map for mGluR5 PAM Benzoxazapine

increase  neutral  decrease

6 December 2011  © Jens Meiler
BCL::PHARMMAP versus CoMFA

<table>
<thead>
<tr>
<th>PHARMMAP</th>
<th>CoMFA</th>
</tr>
</thead>
<tbody>
<tr>
<td>150,000 # Compounds &amp; biological activities used</td>
<td>118</td>
</tr>
<tr>
<td>s-m</td>
<td>Runtime</td>
</tr>
<tr>
<td>NO</td>
<td>Superimposition on common scaffold required?</td>
</tr>
<tr>
<td>Charge Bulk Polarizability # H-bond D/A</td>
<td>Physicochemical properties considered</td>
</tr>
</tbody>
</table>

6 December 2011 © Jens Meiler
Prioritizing Compounds for Chemical Synthesis using 3D SAR

Efficiency: 7.0
EC50 [μM]: 1.33