Automatic Model Generation

Joelle Gola, Olga Obrezanova

19 June 2007
Overview

• Background
• Process
• Techniques
• Illustrative examples
• Next steps
Application of ADME Predictions

Redesign necessary

Virtual library design

Select subset for synthesis chemistries

Select subset for In Vivo testing

Optimise lead

In Silico ADME

In Silico evolution (Model building)

Project knowledge pool

In Vitro ADME

In Vivo ADME
General Process

Data → Model building → Prediction

- Quality Volume
- Descriptors
- Set split
- Techniques
- Validation
- Model selection
- Uncertainty
- Chemical space
- Interpretation
Objectives

- Simplify and speed up the model building process
- Understand and make use of the model output
  - Compound in, prediction out....
- Flexible approach
  - Expert vs. non expert
- Facilitate integration of the model with the decision making process
  - Chemical space analysis
  - Probabilistic scoring
Process

- Checking input data
Process

Structure + Activity

Descriptor Calculation

Feature Selection

Set Split

Modeling Techniques

Model Selection

Prediction

• Whole molecule properties

• ~330 2D SMARTS* based descriptors

• Input own descriptors (additional columns)

• Bespoke 2D SMARTS* based descriptors
Process

Structure + Activity

Descriptor Calculation

Feature Selection

Set Split

Modeling Techniques

Model Selection

Prediction

• Exclude descriptors with:
  - Less than 4% occurrences
  - Standard deviation: 0.0005

• Pair-wise correlated descriptors:
  - Correlation >=95%
Process

Structure + Activity

Descriptor Calculation

Feature Selection

Set Split

Modeling Techniques

Model Selection

Prediction

- Require 3 sets:
  - Training: Building models
  - Validation: Selecting best model
  - Test: Checking best model

- Clustering based method:
  - Structural fingerprint
  - Tanimoto level: 0-1
  - % compounds in the training set

- Input three sets
Process

Structure + Activity

Descriptor Calculation

Feature Selection

Set Split

Modeling Techniques

Model Selection

Prediction

- Classification models:
  - Decision Tree

- Continuous models:
  - Partial Least Squares
  - Gaussian Processes
  - Radial Basis Functions

- Automatically run appropriate models

- User selects preferred techniques
Process

Structure + Activity

Descriptor Calculation

Feature Selection

Set Split

Modeling Techniques

Model Selection

Prediction

- Compare statistical results on validation set:
  - Set of rules to automatically select the best model
  - User choice
Process

- Structure + Activity
- Descriptor Calculation
- Feature Selection
- Set Split
- Modeling Techniques
- Model Selection
- Prediction

- Further validation of the chosen model using test set
- Uncertainty in prediction
- Glowing Molecule
Admensa Process

- **Input data**
  - Descriptor calculation
  - Feature selection
  - Set split

- **Test**
- **Validation**
- **Training**
  - **Feature selection**

- **Model building**
- **Feature selection**
- **Model selection**
- **Model prediction**
- **Save model**

- **Minimum user input**
Techniques: Classification Models

- Recursive partitioning approach to building classification models
- Based on C4.5 approach developed by Ross Quinlan

Rules:

- Generate up to 20 different models by automatically varying method settings:
  - Simple decision trees with various stopping conditions
  - Pruned decision trees
  - Rule sets built from decision trees
Techniques: Continuous Models

- Suite of modeling techniques include:
  - Partial Least Squares
  - Gaussian Processes
  - Radial Basis Functions
Techniques: Gaussian Processes

- Powerful machine learning technique based on a Bayesian statistical approach
  - Bayes Theorem: 
    \[ P(f \mid Y, X) \propto P(Y \mid f, X) P(f) \]
    
    Posterior distribution \quad Prior distribution
  - Gaussian process defines the distribution over functions
Techniques: Gaussian Processes

- Automatic determination of the model parameters:
  - Inherent ability to select relevant descriptors

- Implemented 5 techniques*:

<table>
<thead>
<tr>
<th>Techniques</th>
<th>Descriptor Selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>GP-Fixed</td>
<td>-</td>
</tr>
<tr>
<td>GP-2DSearch</td>
<td>-</td>
</tr>
<tr>
<td>GP-FVS</td>
<td>√</td>
</tr>
<tr>
<td>GP-RFVS</td>
<td>√</td>
</tr>
<tr>
<td>GP-OPT</td>
<td>√</td>
</tr>
</tbody>
</table>

* Obrezanova et al., J.Chem.Inf.Model., accepted

Increasing computational demand
Techniques: Radial Basis Functions

- For given Y and x, RBF is used to find function f(x)
  - f(x) = Y + noise

- f(x) chosen as:
  \[
f(x) = \sum_{i=1}^{N} a_i \left\| x - x^{(i)} \right\|
  \]

- RBF requires f(x) to pass through all training points:
  \[
y_j = \sum_{i=1}^{N} a_i \left\| x^{(j)} - x^{(i)} \right\|, \quad j = 1...N
  \]

weights \(a_i\) can be found from this linear system of equations
Techniques: Radial Basis Functions

- Good method for small or large data sets
  - But sensitive to noise created by excessive descriptors

Solution:

- Coupled with a genetic algorithm, GA-RBF

- Automatically select RBF or GA-RBF:

<table>
<thead>
<tr>
<th># Compounds per descriptor</th>
<th>GA-RBF</th>
<th>RBF</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 5</td>
<td>√</td>
<td>-</td>
</tr>
<tr>
<td>≥ 5</td>
<td>-</td>
<td>√</td>
</tr>
</tbody>
</table>
Techniques: Performance Measures

- **Classification models**

  \[
  \text{Accuracy} = \frac{\text{Correctly predicted compounds}}{\text{Total number of compounds}}
  \]

  \[
  \text{Kappa} = \frac{(\text{Observed agreement} - \text{Chance agreement})}{(\text{Total} - \text{Observed agreement})}
  \]

- **Continuous models**

  \[
  R^2 = 1 - \frac{\sum_i (y_i^{\text{pred}} - y_i^{\text{obs}})^2}{\sum_i (y_i^{\text{obs}} - y_i^{\text{obs}})^2}
  \]

  \[
  \text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i^{\text{pred}} - y_i^{\text{obs}})^2}
  \]
Techniques: Choosing Best Model

- Selection of the best model is done by its performance on the validation set.

Rules:

- Classification model
  - The higher Kappa statistic (0 < Kappa < 1)
  - Best statistical results on training set

- Continuous model
  - Smaller RMSE
  - Best statistical results on training set
The position of compounds relative to the chemical space of the model is reflected in the reported standard deviation (SD).
Techniques: Interpretation

• Molecule in, prediction out...

“Why did this model predict that value for my molecule?”

“What can I do to my molecule to improve the prediction?”

• Models already encode some of these answers!

- Reveal non-linear as well as linear relationships between property being modelled and the selected descriptors

- Proprietary algorithm: Glowing Molecule
Techniques: Glowing Molecules*

Piperacetazine
logP = 3.9

Derivative 1
logP = 3.0

Derivative 2
logP = 5.0

* Patent pending
Illustrative Examples

• Example 1: BBB classification model
• Example 2: Continuous hERG model
• Example 3: Building a local model
Example 1: BBB Classification Model

- BBB± data set from a CNS library published by Zhao et al. Originally prepared by Adenot and Lahana
  - 1593 compounds
  - Chemically Diverse Set

![Graph showing distribution of BBB+ and BBB-]

- Peptide-like
- Aromatic drug-like
- Steroids
- Small organic compounds
## Example 1: BBB Classification Model

<table>
<thead>
<tr>
<th>Models</th>
<th># Descriptors</th>
<th>Kappa</th>
<th>Accuracy</th>
<th>Training (1273 cpds-112 Descriptors)</th>
<th>Validation (158 cpds)</th>
<th>Test (160 cpds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DT1</td>
<td>7</td>
<td>0.78</td>
<td>0.94</td>
<td></td>
<td>0.88</td>
<td>0.96</td>
</tr>
<tr>
<td>DT2</td>
<td>5</td>
<td>0.77</td>
<td>0.94</td>
<td></td>
<td>0.88</td>
<td>0.96</td>
</tr>
<tr>
<td>DT4</td>
<td>6</td>
<td>0.77</td>
<td>0.94</td>
<td></td>
<td>0.88</td>
<td>0.96</td>
</tr>
<tr>
<td>DT7</td>
<td>1</td>
<td>0.63</td>
<td>0.91</td>
<td></td>
<td>0.70</td>
<td>0.90</td>
</tr>
<tr>
<td>DT13</td>
<td>13</td>
<td>0.87</td>
<td>0.96</td>
<td></td>
<td>0.91</td>
<td>0.97</td>
</tr>
<tr>
<td>DT14</td>
<td>15</td>
<td>0.89</td>
<td>0.97</td>
<td></td>
<td>0.90</td>
<td>0.96</td>
</tr>
<tr>
<td>DT15</td>
<td>15</td>
<td>0.86</td>
<td>0.96</td>
<td></td>
<td>0.95</td>
<td>0.98</td>
</tr>
<tr>
<td>DT16</td>
<td>20</td>
<td>0.84</td>
<td>0.96</td>
<td></td>
<td>0.91</td>
<td>0.97</td>
</tr>
<tr>
<td>DT17</td>
<td>13</td>
<td>0.84</td>
<td>0.96</td>
<td></td>
<td>0.90</td>
<td>0.96</td>
</tr>
<tr>
<td>DT19</td>
<td>28</td>
<td>0.88</td>
<td>0.97</td>
<td></td>
<td>0.92</td>
<td>0.97</td>
</tr>
<tr>
<td>DT20</td>
<td>26</td>
<td>0.86</td>
<td>0.96</td>
<td></td>
<td>0.91</td>
<td>0.97</td>
</tr>
</tbody>
</table>
### Example 1: BBB Classification Model

<table>
<thead>
<tr>
<th>Models</th>
<th># Descriptors</th>
<th>Kappa</th>
<th>Accuracy</th>
<th>Kappa</th>
<th>Accuracy</th>
<th>Kappa</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>DT1</td>
<td>7</td>
<td>0.78</td>
<td>0.94</td>
<td>0.88</td>
<td>0.96</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DT2</td>
<td>5</td>
<td>0.77</td>
<td>0.94</td>
<td>0.88</td>
<td>0.96</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DT4</td>
<td>6</td>
<td>0.77</td>
<td>0.94</td>
<td>0.88</td>
<td>0.96</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DT7</td>
<td>1</td>
<td>0.63</td>
<td>0.91</td>
<td>0.70</td>
<td>0.90</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DT13</td>
<td>13</td>
<td>0.87</td>
<td>0.96</td>
<td>0.91</td>
<td>0.97</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DT14</td>
<td>15</td>
<td>0.89</td>
<td>0.97</td>
<td>0.90</td>
<td>0.96</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DT15</td>
<td>15</td>
<td>0.86</td>
<td>0.96</td>
<td>0.95</td>
<td>0.98</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DT16</td>
<td>20</td>
<td>0.84</td>
<td>0.96</td>
<td>0.91</td>
<td>0.97</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DT17</td>
<td>13</td>
<td>0.84</td>
<td>0.96</td>
<td>0.90</td>
<td>0.96</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DT19</td>
<td>28</td>
<td>0.88</td>
<td>0.97</td>
<td>0.92</td>
<td>0.97</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DT20</td>
<td>26</td>
<td>0.86</td>
<td>0.96</td>
<td>0.91</td>
<td>0.97</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Example 1: BBB Classification Model

<table>
<thead>
<tr>
<th>Models</th>
<th># Descriptors</th>
<th>Kappa</th>
<th>Accuracy</th>
<th>Kappa</th>
<th>Accuracy</th>
<th>Kappa</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training (1273 cpds-112 Descriptors)</td>
<td>Validation (158 cpds)</td>
<td>Test (160 cpds)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DT1</td>
<td>7</td>
<td>0.78</td>
<td>0.94</td>
<td>0.88</td>
<td>0.96</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DT2</td>
<td>5</td>
<td>0.77</td>
<td>0.94</td>
<td>0.88</td>
<td>0.96</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DT4</td>
<td>6</td>
<td>0.77</td>
<td>0.94</td>
<td>0.88</td>
<td>0.96</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DT7</td>
<td>1</td>
<td>0.63</td>
<td>0.91</td>
<td>0.70</td>
<td>0.90</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DT13</td>
<td>13</td>
<td>0.87</td>
<td>0.96</td>
<td>0.91</td>
<td>0.97</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DT14</td>
<td>15</td>
<td>0.89</td>
<td>0.97</td>
<td>0.90</td>
<td>0.96</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DT15</td>
<td>15</td>
<td>0.86</td>
<td>0.96</td>
<td>0.95</td>
<td>0.98</td>
<td>0.95</td>
<td>0.98</td>
</tr>
<tr>
<td>DT16</td>
<td>20</td>
<td>0.84</td>
<td>0.96</td>
<td>0.91</td>
<td>0.97</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DT17</td>
<td>13</td>
<td>0.84</td>
<td>0.96</td>
<td>0.90</td>
<td>0.96</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DT19</td>
<td>28</td>
<td>0.88</td>
<td>0.97</td>
<td>0.92</td>
<td>0.97</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DT20</td>
<td>26</td>
<td>0.86</td>
<td>0.96</td>
<td>0.91</td>
<td>0.97</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Example 1: BBB Classification Model

- DT15 is a rule based model with 10 rules

Rule 1:

If PSA > 108.8 and q137 > 26 and calc-logP <= 3.5 then BBB -
confidence = 0.98

Rule 2:

If thioEther = 0 and PSA <= 128.2 and ed70 = 0 and Negative Charge = 0 then BBB+ confidence = 0.97
Example 2: Continuous hERG Model

- 177 pIC$_{50}$ values from the literature
  - Patch clamp measurements in mammalian cells – mainly HEK293

<table>
<thead>
<tr>
<th>Models</th>
<th># Descriptors</th>
<th>Training (124 cpds-143 descriptors)</th>
<th>Validation (27 cpds)</th>
<th>Test (26 cpds)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$R^2$</td>
<td>RMSE</td>
<td>$R^2$</td>
</tr>
<tr>
<td>PLS</td>
<td>143</td>
<td>0.69</td>
<td>0.770</td>
<td>0.61</td>
</tr>
<tr>
<td>GPFixed</td>
<td>143</td>
<td>0.85</td>
<td>0.538</td>
<td>0.66</td>
</tr>
<tr>
<td>GP2DSearch</td>
<td>143</td>
<td>0.83</td>
<td>0.576</td>
<td>0.65</td>
</tr>
<tr>
<td>GPFVS</td>
<td>41</td>
<td>0.78</td>
<td>0.661</td>
<td>0.68</td>
</tr>
<tr>
<td>GPRFVS</td>
<td>19</td>
<td>0.81</td>
<td>0.611</td>
<td>0.65</td>
</tr>
<tr>
<td>GPOPT</td>
<td>42</td>
<td>0.84</td>
<td>0.558</td>
<td>0.63</td>
</tr>
<tr>
<td>GA-RBF</td>
<td>23</td>
<td>1</td>
<td>0</td>
<td>0.64</td>
</tr>
</tbody>
</table>
Example 2: Continuous hERG Model

Validation set:
$R^2 = 0.68$

Test set:
$R^2 = 0.71$
Example 2: Continuous hERG Model

- Important chemical features for hERG binding:
  - Lipophilicity
  - Negative charge
  - Positively charged nitrogen at pH 7.4
  - Aromaticity index

- Redesign using Glowing Molecule
Example 3: Building a Local Model

- Austin et al. measured logD values at pH 7.4 for 78 compounds

- QSAR model highlighted the importance of lipophilicity and ionization in controlling beta(2) duration

- Wants to predict logD7.4 values for new compounds
Example 3: Building a Local Model

- Current global model was built on 1044 compounds
- logD values measured at pH 7.4 were extracted from Starlite™

<table>
<thead>
<tr>
<th>Models</th>
<th># Descriptors</th>
<th>Training (731 cpds – 176 descriptors)</th>
<th>Validation (154 cpds)</th>
<th>Test (159 cpds)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>R²</td>
<td>RMSE</td>
<td>R²</td>
</tr>
<tr>
<td>PLS</td>
<td>176</td>
<td>0.80</td>
<td>0.85</td>
<td>0.75</td>
</tr>
<tr>
<td>GPFixed</td>
<td>176</td>
<td>0.91</td>
<td>0.57</td>
<td>0.81</td>
</tr>
<tr>
<td>GP2DSearch</td>
<td>176</td>
<td>0.91</td>
<td>0.57</td>
<td>0.81</td>
</tr>
<tr>
<td>GPRFVS</td>
<td>52</td>
<td>0.91</td>
<td>0.57</td>
<td>0.82</td>
</tr>
</tbody>
</table>
Example 3: Building a Local Model
Example 3: Building a Local Model

$R^2 = -1.10$
RMSE = 1.1
### Example 3: Building a Local Model

<table>
<thead>
<tr>
<th>Models</th>
<th># Descriptors</th>
<th>Training (54 cpds – 72 descriptors)</th>
<th>Validation (12 cpds)</th>
<th>Test (12 cpds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLS</td>
<td>72</td>
<td>0.44 0.54</td>
<td>0.37 0.57</td>
<td></td>
</tr>
<tr>
<td>GPFixed</td>
<td>72</td>
<td>0.70 0.39</td>
<td>0.32 0.60</td>
<td></td>
</tr>
<tr>
<td>GP2DSearch</td>
<td>72</td>
<td>0.86 0.26</td>
<td>0.39 0.57</td>
<td></td>
</tr>
<tr>
<td>GPFVS</td>
<td>22</td>
<td>0.82 0.31</td>
<td>0.55 0.49</td>
<td>0.62 0.45</td>
</tr>
<tr>
<td>GPRFVS</td>
<td>6</td>
<td>0.81 0.32</td>
<td>0.46 0.54</td>
<td></td>
</tr>
<tr>
<td>GPOPT</td>
<td>18</td>
<td>0.87 0.26</td>
<td>0.47 0.53</td>
<td></td>
</tr>
<tr>
<td>GA-RBF</td>
<td>7</td>
<td>1 0</td>
<td>-0.13 0.78</td>
<td></td>
</tr>
</tbody>
</table>
Example 3: Building a Local Model

Validation set:
$R^2 = 0.55$
RMSE = 0.49

Test set:
$R^2 = 0.62$
RMSE = 0.45
Next Steps

• New modeling techniques:
  ➢ Apply Gaussian Processes to categorical problems

• New descriptors:
  ➢ Automatically designed descriptors to take into account of features not present in current set of descriptors

• Set split:
  ➢ Look at other techniques
General Process

Data → Model building → Prediction

Quality Volume

Descriptors
Set split
Techniques
Validation
Model selection

Uncertainty
Chemical space
Interpretation
Acknowledgments

- Matthew Segall
- Edmund Champness
- Olga Obrezanova
- Christopher Leeding
- Andre Kramer

admensa-support@glpg.com