QSAR Workbench: Guided QSAR Model Building for nonExperts

Chris Luscombe
UKQSAR Autumn 2011
Talk Outline

- Demands of building QSARs
- Needs of Medicinal Chemistry programmes
- Introducing the QSAR Workbench
- Comments and Discussion

Development of local and global QSAR models remains a key requirement for many drug discovery programs; however increasing pressure on resources means that experts in specific statistical tools are not always available. In this presentation we describe the QSAR Workbench, a collaborative development between GlaxoSmithKline and Accelrys that enables specialists to record and publish best practices through an easy to use graphical interface. The recorded workflows can be replayed by both QSAR expert and non-QSAR expert against new datasets. The workbench can be used to explore the massive potential model space defined though the combinatorial use of different statistical tools, descriptors-sets and training/test-set splits. The use of Pipeline Pilot to build the backend framework means that the system is easily extensible in terms of available statistical tools, descriptors and model analysis methods.
Hello World,
We are from *Insilico*, take me to your President

*paraphrased from Dennis Smith, DMPK, Pfizer R&D
DDT Vol. 7, No. 21 November 2002*

- Computational methods for the drug discovery and development process tend to exist in a parallel universe in a different time zone.

- ‘No single *modelling approach* can be used to predict the full range of ADME properties that are desired’.

- ‘It is relatively simple to develop models that fit an entire data set, but typically, such models do not predict new data sets well’.

- we have already established that conventional screening cannot stretch to providing the ADME information and guidance in a cost effective and timely manner, leaving the stage free for the *insilicoids* to rescue the ‘real world’.

- The ‘real world’ needs ‘qsar’ like never before…
Cornerstones of QSAR modelling

Data
• Quality
• Amount
• Coverage
• Train/Test

Expertise
Resource
Best Practice
Validation

Descriptors
• Macro properties
• Structural
• 2D or 3D
• Relevance

Techniques
• Wide Choice
• Interpretative or Blackbox
• Categorical or Continuous
• Software
Automation Drivers in QSAR

Maximise impact of QSAR experts

Science
- Temporal performance
- Descriptors
- Methods
- Data preparation
- Domain of applicability

Technology
- Business
  - Predictive performance
  - Multiple endpoints
  - New data
  - Right model right time
  - Models in chemistry design workflow

Science and Technology are connected, highlighting the importance of integrating scientific methods with technological advancements to enhance the impact of QSAR experts.
QSAR and Drug Discovery process

Library design - Hit triage - Leads - Candidates

Filters/alerts

Global models (statistical, probabilistic)

Local/project models (mechanistic, knowledge)

[First generation models] [Second generation models]
An analysis of the crucial ADME processes for which predictive models are available or are being developed. This figure does not suggest a logical flow in ADME studies, but rather tries to group the problem areas for which predictive models could be of help.
The Promise of Automation

- Automates the routine work so that modelers are more productive

- Moves modelers focus from building models to interpreting results
  - Why did this model outperform all others?
  - Are there combinations of models that cover molecular space better?
  - Are there descriptors that always work well/poorly? Why?
  - What are the time based performance characteristics?

- Push acceptable models to deployment
  - Models web-service

- Better understanding of model spaces
  - Can we be better about knowing (quantifying) the reliability of our predictions?
System Architecture

- **Application Services**
- **Process Management**
  - **Response Provider Adapters**
  - **Asynchronous Task Runner**
  - **Process Builders**
    - **Project Validator**
    - **Descriptor Provider Adapters**
  - **Distributed Job Manager**
  - **Descriptor Provider Services**
- **Flat Files**
- **Data Marts Kate**
- **Compound Dimension**
- **CV Brute Param. Simple**
- **System State**
- **Final Model DB**
- **Blades Desktops**
Rapidly develop, validate and deploy QSAR models

Developed in a collaboration between the Accelrys Professional Services team and pharmaceutical company GlaxoSmithKline, the Accelrys QSAR Workbench is a commercially available, web-based solution that automates and accelerates the development, validation and deployment of predictive Quantitative Structure-Activity Relationship (QSAR) models.

Built on the Pipeline Pilot platform, the QSAR Workbench utilizes native QSAR methods and easily integrates with other statistical tools—helping experts and non-experts alike save time, reduce costs, collaborate more effectively and speed research by leveraging robust, predictive models. The QSAR Workbench reduces modeling time from days to hours and enables chemists to make faster, better decisions.
QSAR Workbench: Automating the Expert

A framework to run each step

Can save and replay

ROC Curve for Model 1 (Accuracy 0.959: Excellent)

Confusion Matrices

Model Predictions for Test Set

Statistics

<table>
<thead>
<tr>
<th></th>
<th>Training</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlation Coefficient</td>
<td>0.74</td>
<td>0.75</td>
</tr>
<tr>
<td>Determination Coefficient</td>
<td>0.55</td>
<td>0.56</td>
</tr>
<tr>
<td>Bias</td>
<td>0.00</td>
<td>-0.22</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.53</td>
<td>0.60</td>
</tr>
<tr>
<td>SCE</td>
<td>52.88</td>
<td>67.58</td>
</tr>
<tr>
<td>Size</td>
<td>187</td>
<td>188</td>
</tr>
</tbody>
</table>
“Automated” builder according to preconfigured expert settings
Build Models Button
Model Triage I – Browse Models Individually
Model Triage II – Compare Models of interest

Interactive plots

Select subsets of models
Key Automated feature choices – Train and Test set splitting choices

- Random:
- Diverse Molecules:
- Random Per Cluster:
- Individual Clusters (Optimised):
- Individual Clusters (Random)

What ratio of A/B should we choose?
Key Automated feature choices – Descriptor subset selections and combinations

- SciTegic Descriptors
  - AlogP Group
  - Estate Keys
  - Fingerprints
  - Molecular Properties2D
  - Molecular Property Counts
  - Surface Area and Volume2D
  - Topological Descriptors

- GSK Descriptors
  - In house developed sets

Prozac
Key Automated feature choices – Modelling Techniques

- Least-Squares
  - Partial-Least-Squares (PLS)
  - Random Forest Regression
  - Genetic Function Approximation (GFA)

- k-Nearest-Neighbor (kNN)
  - R Neural Network (NN)
  - R Support Vector Machine (SVM)

- Linear Discriminant Analysis (LDA)
  - Recursive Partitioning (RP) tree
  - Bayesian Classifier
  - RP Forest

Continuous Response

Categorical Response
Expert Mode – Access to all the parameters

Publish a method (a series of tasks executed in a certain order) using the form below. Once a method is published, other users can replicate the steps taken to build models in this project. Individual task options can be viewed as a tooltip on the Task Name.

<table>
<thead>
<tr>
<th>Task Name</th>
<th>Original Run Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>01 Prepare Chemistry</td>
<td>17 Oct 2011 11:18:44</td>
</tr>
<tr>
<td>01 Split Data</td>
<td>17 Oct 2011 11:18:57</td>
</tr>
<tr>
<td>01 Split Data</td>
<td>17 Oct 2011 11:19:17</td>
</tr>
<tr>
<td>01 Calculate Descriptors</td>
<td>17 Oct 2011 11:20:16</td>
</tr>
<tr>
<td>02 Create Descriptor Subset</td>
<td>17 Oct 2011 11:20:36</td>
</tr>
<tr>
<td>02 Create Descriptor Subset</td>
<td>17 Oct 2011 11:21:22</td>
</tr>
<tr>
<td>02 Create Descriptor Subset</td>
<td>17 Oct 2011 11:21:45</td>
</tr>
<tr>
<td>02 Create Descriptor Subset</td>
<td>17 Oct 2011 11:22:02</td>
</tr>
<tr>
<td>01 Build Models</td>
<td>17 Oct 2011 11:33:18</td>
</tr>
</tbody>
</table>
GSK experience to date

- QSAR models have been generated for Global scenarios
  - In-house solubility model \( n = 16262 \)
  - Whole Blood Binding \( n = 3921 \)
  - hERG patch express \( n = 2124 \)
  - Protein – Protein Inhibition \( n = 1066 \)
  - Cardiovascular ion channels

- And Local models of…..
  - Programme A (Safety assessment related endpoint)
  - Programme B (Selectivity endpoint)
  - Programme C (Potency assay)

- but it’s still early days…. 
Automate the model building process to make it quick and easy to do
Capture ‘best practices’ for model building
Allow chemist to focus on the model validity

Model will be predictive but not necessarily insightful.
Can easily deploy to web service
Will not provide an all singing, all dancing bespoke modelling solution
Will not work everytime
Benefits of this approach

- **Maintainability**
  - Workflows and components
  - Self-contained system
    - Reduced reliance on external groups

- **Extensability**
  - New descriptors
  - New modelling methods

- **Direct integration with existing systems for model publication**
  - Automated model reporting and publication
  - Web-services

- **Encapsulates “expert” knowledge**
  - GSK descriptors
  - GSK model parameterisation

- **Capturing workflows**
  - “Big Green Button”
    - Part or whole of workflow

- **Building and exploring larger model spaces**
Change Management Challenges

**Modellers**
- Trusting in automation
- Letting go of old practices (old favourites)
- Intellectual emphasis shifts from model building to model selection

**Chemists**
- Using multiple predictive models together
  - Multi Objective analysis
  - Not using as filters in XL!
- Using continually evolving models (as more information becomes available)
  - What was correctly predicted last week might not be this week, but there should be fewer false predictions overall…
Figure 4 | **Towards prediction paradise.** As more and more robust models for the crucial endpoints in the drug discovery process become available, we will increasingly be in a position to map out the potential qualities of a new chemical purely from its molecular structure and appropriate descriptors using a suite of predictive models. These range from models for simple physicochemical properties, such as hydrogen bonding-capability, molecular mass, solubility and lipophilicity (log D), to models for ADME properties, such as percentage drug absorbed and bioavailable, clearance, volume of distribution and half-life, to complex endpoints, such as the binding (IC\textsubscript{50}) to the molecular target of the new drug, its required dose and toxicity potential.
Welcome *insilicoids* to the ‘real world, real time zone’; get this right and do it now, and we will make you the President.

- **GSK**
  - Stephen Pickett, Darren Green
  - Chris Keefer, Nate Woody, Chris Bizon (AME)

- **Accelrys**
  - Adrian Stevens, Noj Malcolm, Richard Cox, David Nicolaides
  - Julian Wilmot, Michael Kanz, David Podesta