Automated QSAR Modelling to Guide Drug Design

Olga Obrezanova, Matthew Segall, Joelle Gola and Edmund Champness

> Zing Computational Chemistry Conference March 2009



© Copyright 2009 Galapagos NV

Automatic model generation

- The rapid design-test-redesign cycle of modern drug discovery demands fast model building
- Automatic modelling processes allow
 - > exploring large numbers of modelling approaches efficiently
 - making QSAR model building accessible to non-experts





- Automatic Model Generation process (AMG)
 - Stages of the process
 - > Gaussian Processes modelling techniques
- 'Manual' model versus 'automatic'
 - Blood-brain barrier penetration
 - > Aqueous solubility
- Building QSAR model to guide drug design



Automatic Model Generation Process



© Copyright 2009 Galapagos NV

Automatic model generation



 Splitting data into training, validation and test sets (by cluster analysis)

 Descriptor calculation and filtering (2D SMARTS, logP, TPSA, MW, charge etc.)

 Modelling techniques (PLS, Radial Basis Functions with genetic algorithm, Gaussian Processes, Decision Trees)

 Selection of the best model by performance on the validation set

• Test set is an **independent** set

Automatic model generation



Modelling techniques: Gaussian Processes

- A machine learning method based on Bayesian approach
- Advantages:
 - > Does not require a priori determination of model parameters
 - > Nonlinear relationship modelling
 - > Built-in tool to prevent overtraining no need for cross-validation
 - > Inherent ability to select important descriptors
 - > Provides uncertainty estimate for each prediction
- Sufficiently robust to enable automatic model generation



Modelling techniques: Gaussian Processes

 Define **prior distribution** over functions (controlled by hyperparameters, covariance function – ARD function)

• **Posterior distribution**: retain functions which fit experimental data

• **Prediction** is the mean of posterior distribution.

 Standard deviation of the distribution provides estimate of the uncertainty in prediction





'Automatic' models versus 'manual'

Applications: blood-brain barrier penetration and aqueous solubility



© Copyright 2009 Galapagos NV

Blood-brain barrier penetration (logBB)

- Data set of 151 compounds with logBB values (collected from literature)
- 'Manually' built model (random set split TRN=108, TEST=43)
- Build a model by the automatic model generation (AMG) process (apply to all 151 compounds)
- Compare `automatic' and `manual' models by testing on external data
 - 143 compounds from 'Abraham' set not present in the initial set (Abraham et al. J.Pharm. Sci., 2006, 95)



Blood-brain barrier penetration (logBB)

`Manual' model

> 2D SMARTS descriptors reduced by FVS, various modelling techniques (PLS, RBF, MLR) – performance supervised on test set

Final model is built by Radial Basis Functions on 7 descriptors (logP, flexibility, charge, hydrogen bonding...)

- 'Automatic' model
 - cluster at t=0.7, val=23 comp, test=22 comp
 - Best model by GP with nested sampling
 - 162 descriptors

manual
Test set
R²
0.73
RMSE
0.36



BioFocus DPI



Blood-brain barrier penetration Performance on external 'Abraham' test set

Model	RMSE pred	% pred within ±0.4 log unit	% pred within ±0.8 log unit	R²	r ² _{corr}	RMSE
manual	0.36	62.9	93.0	0.39	0.44	0.44
automatic	0.44	63.6	90.9	0.27	0.36	0.49







Automatic rebuilding logBB model to include new data

• Original 151 compounds and 143 compounds from 'Abraham' set

Best model – GP with 2Dsearch on 167
descriptors:

Set	Ν	R ²	RMSE
TRN	205	0.80	0.29
VAL	44	0.73	0.33
TEST	43	0.67	0.35

• Improvement in prediction of 30 compounds from 'Abraham' set, now in val and test sets:

- current model RMSE=0.27
- previous automatic model RMSE=0.44





[®]Aqueous solubility

- 3313 compounds with intrinsic aqueous solubility (logS, S in μ M)
 - From PHYSPROP database (Syracuse Research Corporation, SRC)
- Automatic model produced by Gaussian Processes with 2D search
- External test data 564 compounds from 'Huuskonen' set

Huuskonen J., J. Chem. Inf. Comput. Sci., 2002, 42

manual





BioFocus DPI



Aqueous solubility Performance on external 'Huuskonen' test set

Model	Desc	% pred within ±0.7 log unit	% pred within ±1.4 log unit	R ²	r² _{corr}	RMSE
manual	108	39.9	70.9	0.68	0.80	1.28
automatic	166	54.1	85.9	0.82	0.86	0.96





Building QSAR model to guide drug design

Case Study



© Copyright 2009 Galapagos NV

Building QSAR model to guide drug design

 The automatic model generation algorithm is implemented in the StarDrop environment for decision support in drug discovery and is referred to as the Auto-Modeler

 QSAR models can be used to predict new compounds together with the Glowing Molecule visualisation tool

• Interpret SAR and guide redesign of compounds to overcome liabilities





The 'Glowing Molecule': visualisation tool

 Makes a link between predicted property and compound's structure

- "Why is a property value predicted?"
- >"Where can I change this property?"
- ≻Interpret SAR
- ➢Guide efficient redesign of molecules
- No-more 'black box' models!



logP property

BioFocus DPI



Building QSAR model to guide drug design

- QSAR model for Target X affinity
 - > 138 compounds with pKi data from screening against 'Target X'
 - Apply Auto-Modeler

The best QSAR model of affinity:

Set	R ²	RMSE
VAL	0.96	0.23
TEST	0.95	0.29

BioFocus

A **Galáp**agos Company

- Predicting affinity
 - > Additional experimental affinity data for 10 new compounds

> Model predictions correlate very well with the experimental data $R^2=0.98$, RMSE=0.22



Building QSAR model to guide drug design

- Need to have balance of potency and ADME properties, hence incorporate predictions from StarDrop ADME models (logS, hERG, BBB, HIA, PPB, logP, 2C9 affinity, pgp ...)
- Apply probabilistic scoring all compound data integrated to allow prioritization
- Score new molecules against project profile
 - Scoring profile is for an orally bioavailable, potent molecule for a non-CNS target (incorporates desired project criteria and their importance)
- Resulting score estimates each compound's likelihood of success against the project profile



Building QSAR model to guide drug design





BioFocus DPI A Galápagos Company

Building QSAR model to guide drug design





- Described the automatic model generation process for QSAR modelling
- The process was applied to modelling blood-brain barrier penetration and aqueous solubility
- Automatic models compare well to ones built manually. The automatic process is robust, much quicker than manual building and can be applied by non-experts
- The case study demonstrates how building a QSAR model can help to understand SAR for a chemical series and redesign compounds to overcome liabilities



Acknowledgements

- Matthew Segall
- Chris Leeding
- Ed Champness
- Joelle Gola
- Andre Kramer









© Copyright 2009 Galapagos NV

Gaussian Processes: Hyperparameters

- Learning the Gaussian Process ~ finding hyperparameters
 - Optimize the marginal log-likelihood (prevents overtraning, no need for validation set)
- Techniques for finding hyperparameters
 - "Fixed" values for length scales. Search for noise parameter
 - Forward variable selection provides feature selection
 - > Optimization by conjugate gradient methods
 - Length scales show which descriptors are most relevant
 - Nested sampling
 - Search in the full hyperparameter space
 - Search does not get trapped in local maxima



BioFocus

Modelling techniques. Gaussian

Processes



- Learning the Gaussian Process ~ finding hyperparameters
 - Length scale parameters (one for each descriptor) → identification of relevant descriptors
 - > Noise parameter
- Find hyperparameters by optimizing the marginal log-likelihood
 - It controls the trade-off between the model complexity and fit to the data no need for cross-validation, prevents overfitting, does not require user intervention

BioFocus



Solution Series Series Gaussification

- Introduce 'latent' variable a(x); model it by GP
- Latent function a(x) can be turned to probability by following transformation

$$P(y=1 \mid x, a(x)) = \frac{1}{1 + \exp(-a(x))}$$

 Obtain two approximations for class probability





Solution Gaussian Processes for binary classification

- We used an approximation method of lower and upper variational bounds (Gibbs and Mackay, 1997)
 - provides approximations for class probability (lower and upper bounds – 2 approximations)
- Learning the classifier
 - Optimization of parameters: hyperparameters (as in GP regression) and variational parameters
 - > Techniques: conjugate gradient optimization, nested sampling
 - Number of parameters ~ (#descriptors) + 2 (#compounds)

