An application of Multi-variate adaptive regression splines (MARS) in QSRR

R. Put*, D.L. Massart, Y. Vander Heyden

Dept. of Pharmaceutical and Biomedical analysis,
Pharmaceutical Institute,
Vrije Universiteit Brussel,
Laarbeeklaan 103, B-1090 Brussel

* E-mail: rafput@vub.ac.be

This poster can be downloaded at: http://put.be.tf/

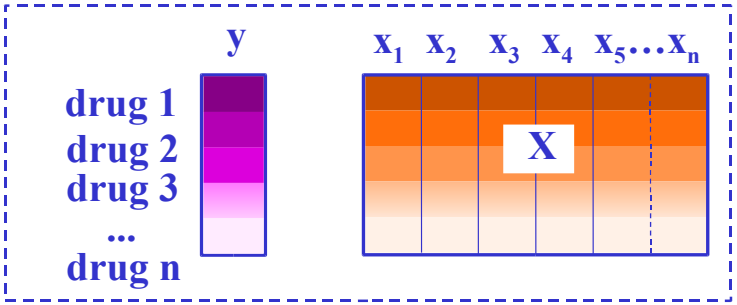
Internet Electronic Conference of Molecular Design 2003

Introduction

- **Retention prediction** for High Performance Liquid Chromatography

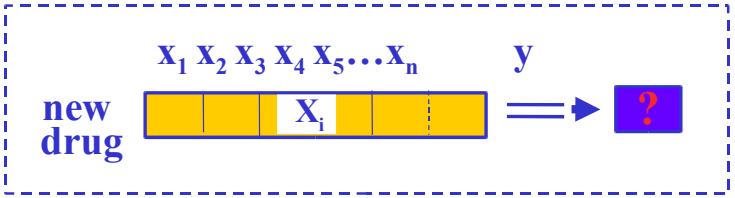
...

Quantitative Structure Retention Relationships (QSRR) [1]



Training set ⇒ QSRR Model

- x_1 … x_n : molecular descriptors
- y : chromatographic retention



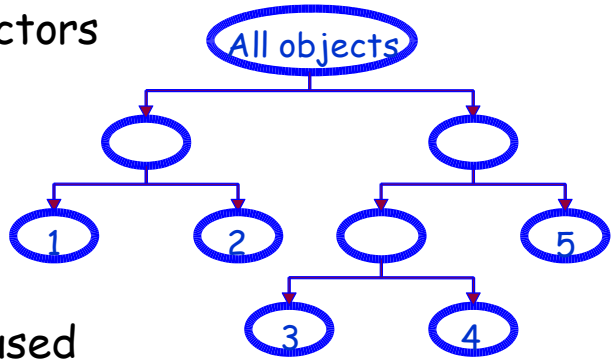
Predict y for new molecules

- Which Molecular descriptors to include in the QSRR-model ?
  - ➡ selection based on chromatographical knowledge (log P,…)
  - ➡ selection of the "best" descriptors
    - feature selection techniques (Genetic Algorithms,…)
    - during the model building (CART, MARS,…)

- Aim : study the use of MARS (and CART for feature selection) in a QSRR context

CART [2]

- **Goal** : modeling the response variable, using independent predictors



- **Splits** : - Defined by 1 predictor
  - Additional primary splits
    ⇒ most important predictors
  - Surrogate splits
    ⇒ used for missing values of the predictors used

- **Result** : - set of predictors is selected in the model
  - Classes with low, intermediate and high response values
  - Mean of the responses within each class = predicted value for new objects

MARS [3]

- Multivariate non-parametric adaptive regression procedure
- Global MARS model : weighted **sum** of all **local models** :

$$\hat{f}_M(\mathbf{x}) = a_0 + \sum_{m=1}^{M} a_m B_m(\mathbf{x})$$

a_0 :	coefficient (constant basis function)
$B_m(\mathbf{x})$:	mth basis function
a_m :	coefficient of the basis function
M :	number of basis functions included

MARS

- **3 steps** in the model building:

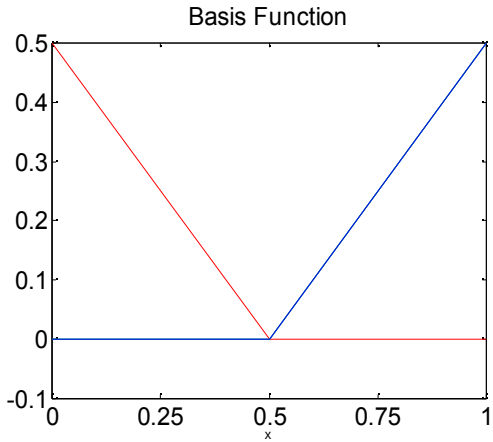
① **Constructive phase**

- Similar to recursive partitioning (CART)
- Introduces local models in several (overlapping) regions of the space of possible predictors :

  ⇒ defined as basis functions = | - one single spline function
  | - the product of 2 (or more) splines (interaction different predictors)

- Splines

$$(x-t_0)_+ = \begin{cases} (x-t_0), & \text{if } x > t_0, \\ 0, & \text{otherwise} \end{cases} \qquad (t_0-x)_+ = \begin{cases} (t_0-x), & \text{if } x < t_0, \\ 0, & \text{otherwise} \end{cases}$$



- ⇒ Overfitted MARS model

② **Pruning phase**

- Backward elemination procedure : some basis functions are deleted
- The generalized cross validation criterion is used :
- ⇒ Sequence of smaller and smaller MARS models

$$GCV(M) = \frac{1}{n} \frac{\sum_{m=1}^{n} (y_i - \hat{f}_M(\mathbf{x}_i))^2}{(1 - C(M)/n)^2}$$

$C(M) = M + dc$

M : number of terms

c : number of nonlinear terms

$d = 2$

③ **Selection of the optimal model**

- Using cross-validation (CV) (default: Leave-1-out) or an independent test set

Molecular Representations

- For all molecules the geometrical structure was optimized using **Hyperchem 6.03** Professional software (Hypercube, Gainesville, Florida, USA).
  - ➡ The Polak-Ribiere conjugate gradient algorithm was used for the geometry optimization with a RMS gradient of 0.05 Kcal / (Å mol) as stop criterion.
  - ➡ Energy minimization was done with the Molecular Mechanics Force Field method (MM+).
- The Cartesian coordinates matrix of the positions of the atoms in the molecule, and which is resulting from these 3D representations were used for the calculation of the molecular descriptors [4] using the **Dragon 1.1** software of Todeschini et al. [5] (http://www.disat.unimib.it/chm/Dragon.htm)
- A selection of the **molecular descriptors** was made, in a way that only 0D, 1D, 2D and experimental descriptors were used, derived from the above-mentioned representation. The following groups of descriptors, as defined in Dragon 1.1, were calculated: 56 constitutional descriptors, 69 topological descriptors, 20 molecular walk counts, 21 Galvez topological charge indices, 96 2D autocorrelations and 3 empirical descriptors.
- Additional log P values were obtained from Detroyer et al. using **LOGKOW** [6] (http://esc.syrres.com/interkow/kowdemo.htm)

Methodology

Data

83 basic drugs

Chromatographic retention : log k_w on Unisphere PBD column (polybutadiene-coated alumina) at pH 11.7 using isocratic elutions [4]

266 molecular descriptors [5] :

- log P values : LOGKOW (Detroyer et al.) [6]
- by Dragon 1.1 [7] based on 3D representations optimized in Hyperchem 6.03 (MM+, Polak-Ribière)

- **CART :** decision trees were build using the TreePlus module [8] for S+2000
- **MARS :** - an in-house algorithm based on the original MARS method was used in the Matlab 5.3 environment. (Pruning : GCV was alternated with 20-fold CV)
  - Model selection using leave-one-out CV (default) and Monte Carlo CV (MCCV)

MCCV [9]
1) randomly data (n) → Testset (n_v) / Calibration set (n_c)
2) calculate RMSECV
3) repeat this N times ($N=n^2$)

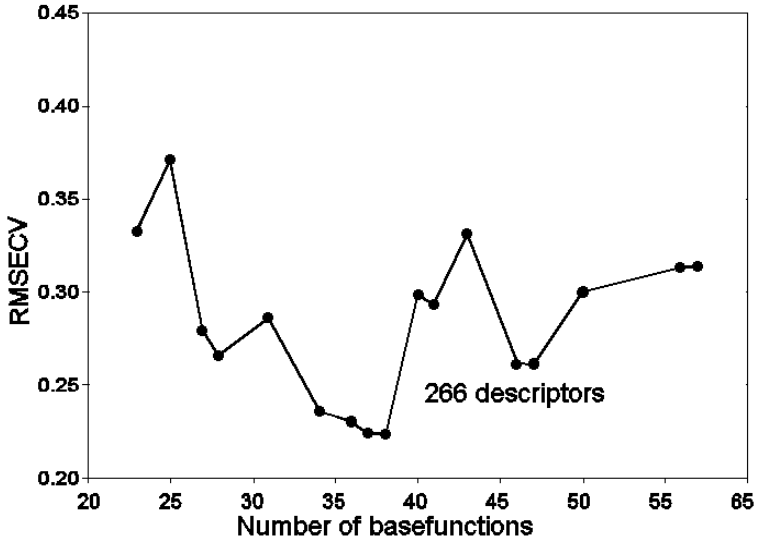
$$RMSECV = \sqrt{\frac{PRESS}{n_v}}$$

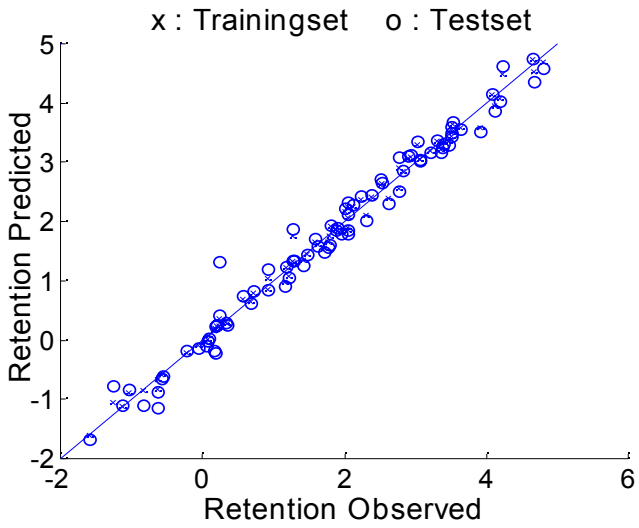
$$PRESS = \sum \left(y_{predicted} - y_{observed} \right)^2$$

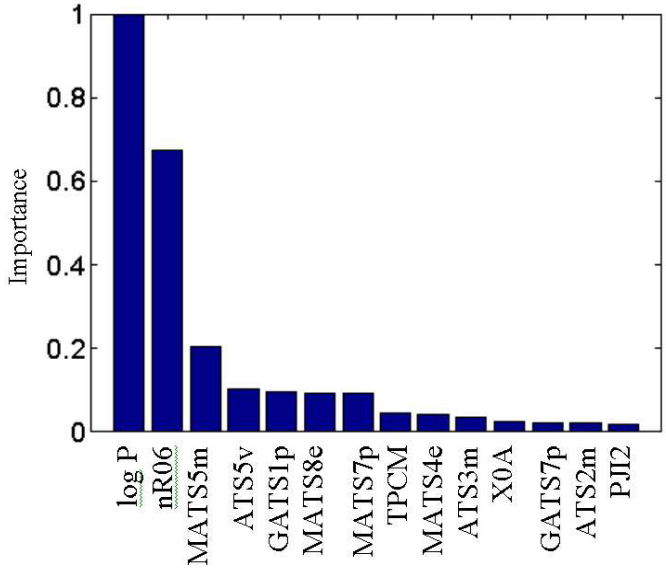
Results & Discussion [10,11]

MARS leave-1-out CV

- **Optimal model :**
  - 34 basis functions
  - $RMSECV_{(leave\text{-}1\text{-}out)}$ = 0.2358

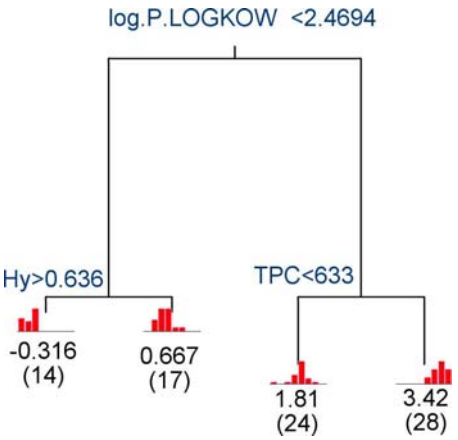






Results & Discussion

CART + MARS leave-1-out CV

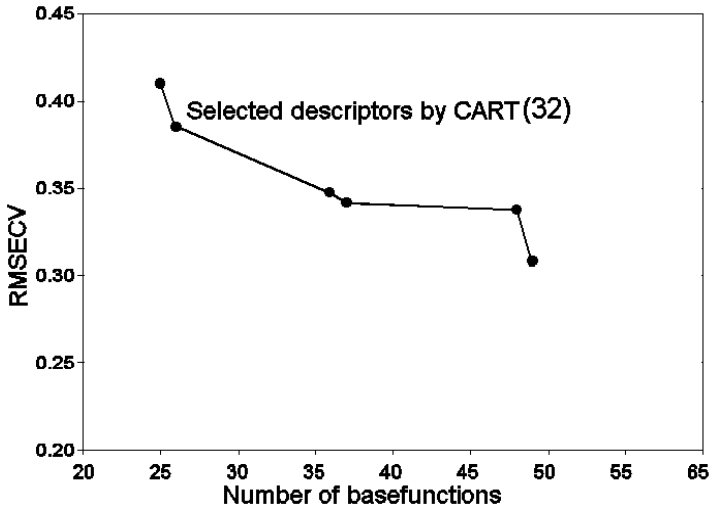


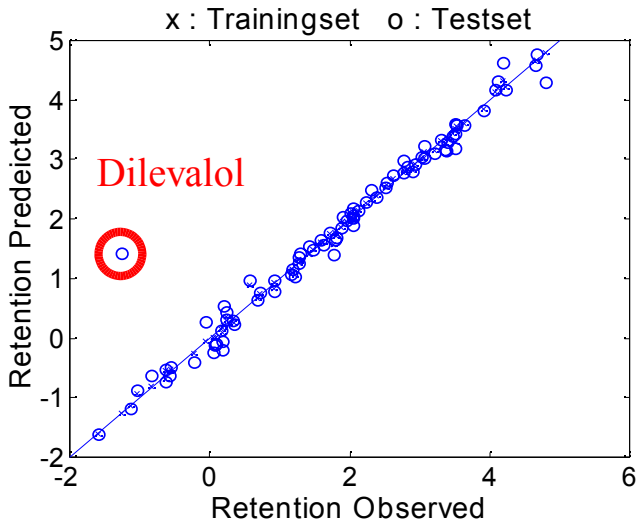
- **CART :** model with 4 leaves -- 3 splits -- 32 molecular descriptors selected (3 + primary + surrogate splits)

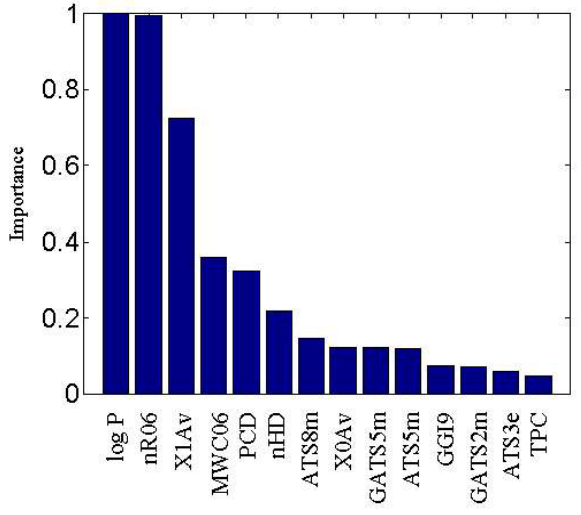
- **Optimal (?!) MARS model :**
  - 49 basis functions (very complex)
  - RMSECV = 0.3373 (leave-1-out CV)
  - !! Dilevalol : very bad prediction

OVERFITTING ??

=> Investigate with MCCV







Results & Discussion

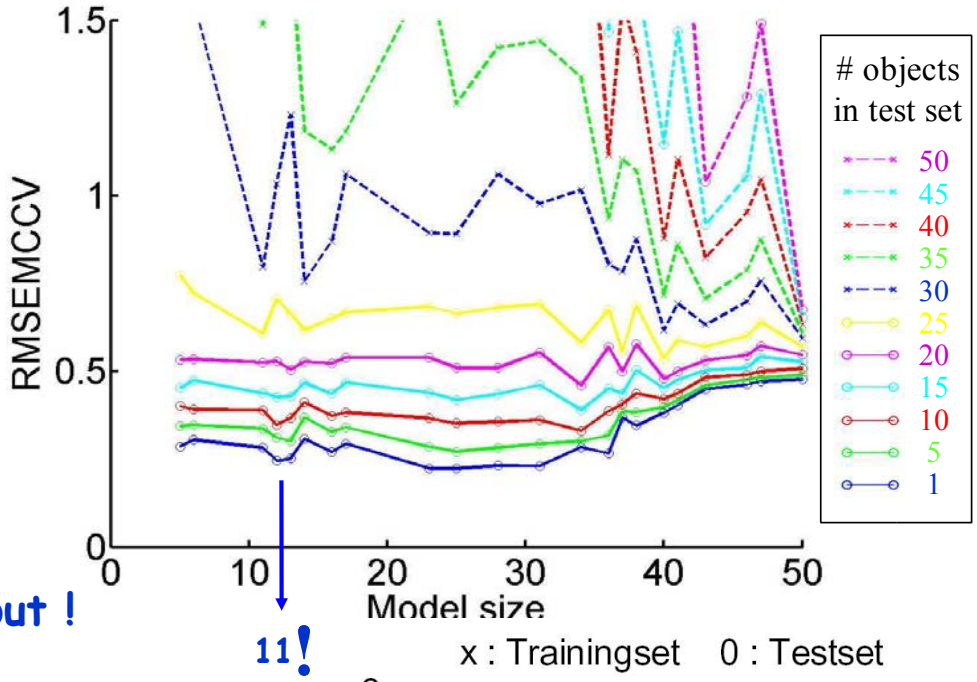
MARS Monte Carlo CV

- **Influence of the CV test set size:**

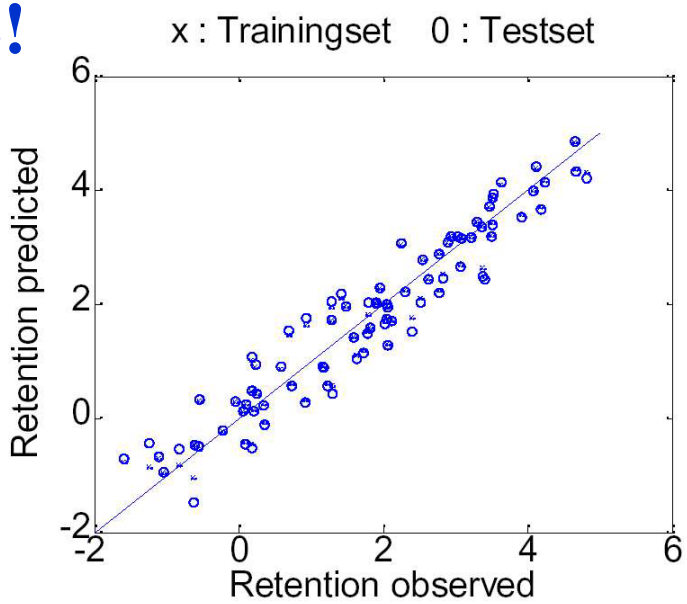
Conclusions :

- Default testset size (50% of objects) (~42 objects) may be too large !
- Optimal model size = 11

<<< Leave-1-out !



- **Optimal model :**
  - 11 basis functions
  - $RMSECV_{leave\text{-}1\text{-}out\ CV}$ = 0.4766



Conclusions

- The MARS methodology shows potential in a QSRR context
  - overall good predictions
  - molecular descriptors used in the model are interpretable

! leave-1-out CV may lead to overfitted MARS models

→ Monte Carlo cross-validation may be preferable

- CART can be used for feature selection prior to MARS

[1] R. Kaliszan, Quantitative Structure-Chromatographic Retention Relationships. Wiley-Interscience, New York, NY, 1987.
[2] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression trees, Monterey, 1984
[3] J. H. Friedman, Multivariate adaptive regression splines, Annals of Statistics 19 (1991) 1-141
[4] A. Nasal, A. Bucinski, L. Bober, R. Kaliszan, Int. J. Pharm. 159 (1997) 43-55
[5] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley-VCH, Weinheim, 2000
[6] A. Detroyer, V. Schoonjans, F. Questier, Y. Vander Heyden, A.P. Borosy, Q. Guo, D.L. Massart, J. Chromatogr. A, 897 (2000) 23–36
[7] R. Todeschini, V. Consonni, Dragon software version 1.1, http://www.disat.unimib.it/chm/Dragon.htm
[8] G. De'Ath. New statistical methods for modeling species-environment relationships, Ph.D. Thesis. James Cook University, Townsville, Australia, 1999
[9] Picard, R. R.; Cook, R. D. J. Amer. Statist. Assoc. 1984, 79, 575-583
[10] R. Put , C. Perrin , F. Questier , D. Coomans , D.L. Massart , Y. Vander Heyden, J . Chromatogr . A 988 (2003) 261–276
[11] Multivariate adaptive regression splines in chromatographic quantitative structure-retention studies, R. Put, Q.S. Xu, D.L. Massart and Y. Vander Heyden, In preparation