THE EFFECTS ON SENSITIVITY OF THE STRUCTURE-ACTIVITY MODELS

Sorana D. BOLBOACĂ[1] and Lorentz JÄNTSCHI[2]

[1]"Iuliu Haţieganu" University of Medicine and Pharmacy Cluj-Napoca, 6 Louis Pasteur, 400349 Cluj-Napoca, Romania, http://sorana.academicdirect.ro
[2]Technical University of Cluj-Napoca, 103-105 Muncii Bvd, 400641 Cluj-Napoca, http://lori.academicdirect.org

ABSTRACT

An analysis to identify the effects of standardized residuals, hat-matrix and Cook's distance on the sensitivity of regression structure-activity models was conducted. A sample of 250 phenolic compounds with measured toxicity on Tetrahymena pyriformis was the input data of the analysis. Simple linear models was identify to estimated toxicity as function of LUMO (energy of the lowest unoccupied molecular orbital) and octanol/water partition coefficient expressed in logarithm scale. To accomplish the aim of the research the compounds identified with values of residual, hat-matrix leverage and Cook's distance higher than thresholds were remove and the model was rebuilt. An assessment of the models was conducted in order to identify which approach is able to identify compounds with significant influence on the QSAR model.

AIM

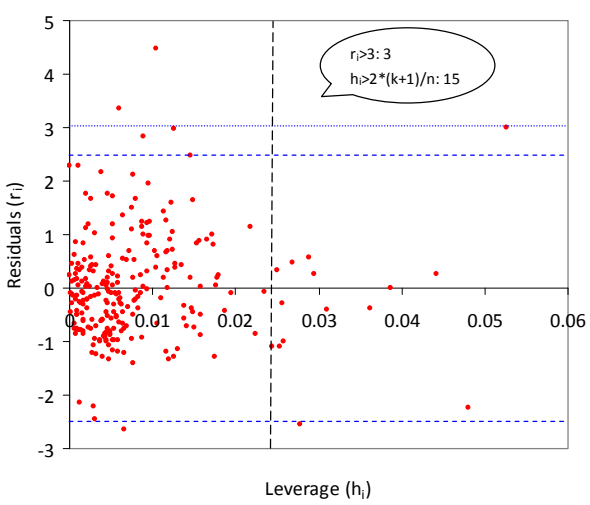
Model sensitivity analysis was conducted to assess the effects on statistical characteristics of linear regression model obtained to link the toxicity with compounds hydrophobicity and energy of the lowest unoccupied molecular orbital.

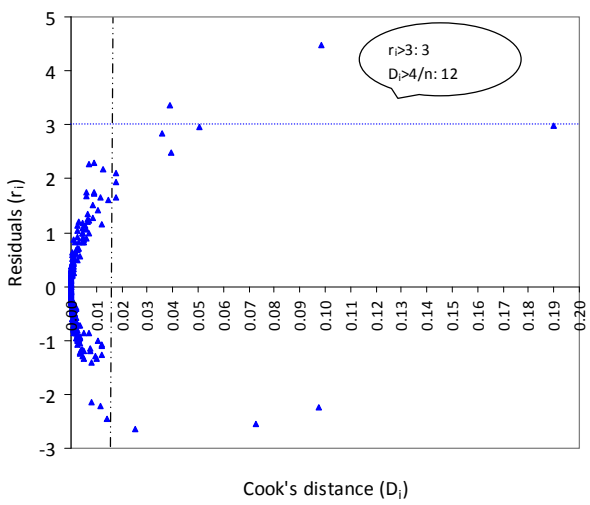
METHODOLOGY

A sample of 250 phenolic compounds with measured toxicity on *Tetrahymena pyriformis*, calculated octanol/water partition coefficient (logP) and LUMO (energy of the lowest unoccupied molecular orbital taken from [1]) were taken from a previous published manuscript [2]. Standardized residuals (values higher than 2.5 [3] and 3 [4]), hat-matrix leverage ($h_i>2*(k+1)/n$, where h_i = leverage for i^{th} compound, k = number of independent variables in the regression model and n = sample size), and Cook's distance ($D_i>0.85$ (k=2) [5]; $D_i>1$ [6] and $D_i>4/n$ [7]). The compounds identified with values of residual, hat-matrix leverage and Cook's distance higher than thresholds were remove and the model was rebuilt. The removal was performing whenever an improvement in correlation coefficient was obtained. Several statistical criteria [8] were used to validate and to compare models. Steiger's test [9] at a significance level of 5% was applied to test if correlation coefficients were statistically different between models.

RESULTS

Residuals - hat-matrix - Cook's distance





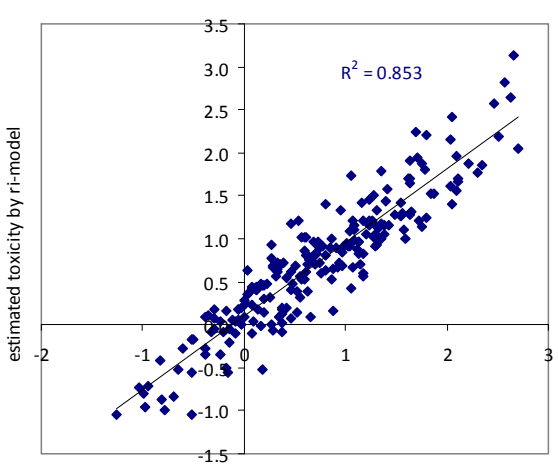
The removal of compounds with hi higher than threshold did not led to any improvement of the model characteristics and was not further investigated.

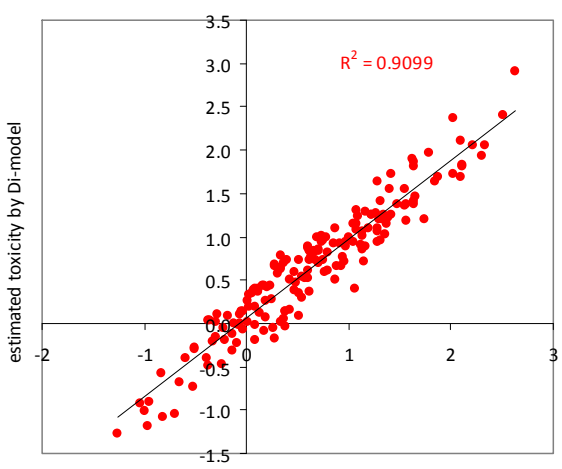
MODELS CHARACTERISTICS

	Full (n=250)	$r_i>3$ (n=219)	$D_i>4/n$ (n=179)
R^2	0.5643	0.8530	**0.9099**
R^2_{adj}	0.5608	0.8517	**0.9089**
F-value	160	627	**889**
p-value	$2.77\cdot10^{-45}$	$1.15\cdot10^{-90}$	$\mathbf{9.94\cdot10^{-93}}$
$R^2_{cv\text{-}loo}$	0.5513	0.8482	**0.9068**
RMSE	0.5491	0.3121	**0.2352**
MAE	0.4051	0.2569	**0.1978**
MAPE	1.5639	0.9754	**0.8280**
SEP	0.5468	0.3107	**0.2338**
REP	73.8294	42.8440	**35.9884**
RMS	0.3002	0.0970	**0.0550**
APV	0.3026	0.0978	**0.0556**
TSE	**3**	**3**	**3**
APMSE	0.0012	0.0004	**0.0003**
%PredErr	37.3708	20.7592	**13.4657**

R^2 = determination coefficient; R^2_{adj} = adjusted determination coefficient
$R^2_{cv\text{-}loo}$ = leave-one-out determination coefficient;
F-value = Fisher's statistics; p-value = probability associated to F-value;
r = residuals; D = Cook's distance; $1 \le i \le n$;
RMSE = root-mean-square error; MAE = mean absolute error;
MAPE = mean absolute percentage error;
SEP = standard error of prediction; REP = relative error of prediction;
RMS = residual mean square; APV = average prediction variance;
TSE = total squared error; APMSE = Average Prediction Mean Squared Error;
%PredErr = percentage prediction error;

RI-MODEL & DI-MODEL





MODELS PERFORMANCES

	Sensitivity	Specificity	Accuracy
Full model	0.9463 [0.9065-0.9698]	0.4000 [0.2702-0.5455]	0.8480 [0.807-0.884]
ri-model	0.9722 [0.9366-0.9881]	0.6923 [0.5358-0.8143]	0.9224 [0.879-0.949]
Di-model	0.9583 [0.9121-0.9808]	0.7429 [0.5793-0.8584]	0.9162 [0.864-0.915]

DETERMINATION COEFFICIENT COMPARISON:

- Full model vs. D_i-Model: $z = 9.1$, $p < 0.0001$
- Full model vs. r_i-Model: $z = 6.8$, $p < 0.0001$
- D_i-Model vs. r_i-Model: $z = 2.6$, $p = 0.0052$

CONCLUSION

Two out of three investigated approaches (withdrawing the compounds with residual influence and of compounds exceeding the Cook's distance threshold) led to increase in both determination and performances. The Cook's distance approach proved the method with both higher determination coefficient and model specificity.

REFERENCES

1. Cronin MTD, Aptula AO, Duffy JC, Netzeva TI, Rowe PH, Valkova IV, Schultz TW. Comparative assessment of methods to develop QSARs for the prediction of the toxicity of phenols to Tetrahymena pyriformis. Chemosphere 2002;49:1201-1221.
2. Zhao YH, Yuan X, Su LM, Qin WC, Abraham MH. Classification of toxicity of phenols to Tetrahymena pyriformis and subsequent derivation of QSARs from hydrophobic, ionization and electronic parameters. Chemosphere 2009;75(7):866-871.
3. Tropsha A, Gramatica P, Gombar VK. The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models, QSAR Comb. Sci. 2003;22(1);69-77.
4. Gramatica P. Principles of QSAR models validation: internal and external, QSAR Comb. Sci. 2007;26(5):694-701.
5. Mc Donald B. A Teaching Note on Cook's Distance – A Guideline. Res. Lett. Inf. Math. Sci. 2002;3:127-128.
6. Cook RD, Weisber S. Residuals and influence in regression. New York: Chapman & Hall, 1982.
7. Bollen KA, Jackman R. Regression diagnostics: An expository treatment of outliers and influential cases. In: J. Fox & J. Scott Long (eds.) Modern Methods of Data Analysis. Newbury Park: Sage, 1990, pp. 257-291.
8. Bolboacă SD, Jäntschi L. The Effect of Leverage and Influential on Structure-Activity Relationships. Submitted.
9. Steiger JH. Tests for comparing elements of a correlation matrix. Psychol. Bull. 1980;87:245-251.

Participation at conference was supported by "Iuliu Haţieganu" University of Medicine and Pharmacy Cluj-Napoca.