APPLICATION OF ROBUST REGRESSION IN QSAR MODELING

Rainer Grohmann[a,b], and Torsten Schindler[a]

[a]Novartis Institutes for BioMedical Research, 1235 Vienna, Austria
[b]Department of Biomelecular Structural Chemistry, University of Vienna, 1030 Vienna, Austria

Partial Least Squares Regression is routinely employed in QSAR/QSPR modeling. A number of studies focus on ruggedizing QSAR modeling in experimental design [1], in variable selection [2][3] and in model validation [4]. Partial Least Squares regression is frequently employed using the SIMPLS [5] and NIPALS [6] algorithms. However, these are sensitive to outliers in the dataset. The resulting regression may be skewed towards bad leverage points, whilst the regression accuracy still seems acceptable.
To circumvent the risk of biased results caused by outliers, robust statistical approaches have been developed. These provide estimates for statistical parameters (e.g. covariance matrix), that are not susceptible to the influence of outliers. Furthermore, outliers can clearly be identified.
Hundreds of molecular descriptors are available to map the biological/chemical interface. To discard redundant information contained in the descriptor-set and improve the model accuracy descriptor selection methods are used [2][3]. Descriptor selection techniques are especially important in PLS modeling since regression models increase their fitting capability on increasing the number of variables, even if these are random and are not related to the problem [7].
Here, we present QSAR/QSPR models for public available datasets, generated by classical and robust PLS regression [8] for a variety of variable selection approaches.
Our aim is to investigate how outlier detection methods and variable selection influence each other and to compare the robustness and predictive power of derived PLS models. From these results we will propose an approach to derive QSAR models, based on a cleaned, homogenized set of compounds, with better predictive performance and justified model validation.

[1] Olsson IM., D-optimal onion designs in statistical molecular design, Chemometrics and Intelligent Laboratory Systems, (2004) 73, 37-46
[2] Tropsha A., et. al., Identification of the Descriptor Pharmacophores Using Variable Selection QSAR: Applications to Database Mining, Current Pharmaceutical Design, (2001) 7, 599-612
[3] Xu L., et. al., Comparison of different methods for variable selection, Analytica Chimica Acta, (2001) 446, 477-483
[4] Tropsha A., et.al., The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models, QSAR Combinatorial Science (2003) 22, 69-77
[5] De Jong S., SIMPLS: an alternative approach to partial least squares regression, Chemometrics and Intelligent Laboratory Systems, (1993) 18, 251-263
[6] Wold H., Estimation of principal components and related models by iterative least squares, in Krishnaiah PR (Ed.), *Multivariate Analysis*, Academic Press, New York, 1966, pp. 391-420
[7] Wold S., Validation of QSAR's, Quantitative Structure-Activity Relationship, (1991) 10, 191-193
[8] Hubert M., et. al., Robust methods for partial least squares regression, Journal of Chemometrics, (2003) 17, 537-549