Chemistry Central Journal

ChemistryCentral

Poster presentation

Validation of predicitve modelling techniques in drug design – influence of test set composition

M Matz*, S Rohrer and K Baumann

Address: Magnus Matz, Institut für Pharmazeutische Chemie, Technische Universität Braunschweig, Beethovenstr. 55, 38106 Braunschweig, Germany

* Corresponding author

from 4th German Conference on Chemoinformatics
Goslar, Germany. 9–11 November 2008

Published: 5 June 2009

Chemistry Central Journal 2009, **3**(Suppl 1):P64 doi:10.1186/1752-153X-3-S1-P64

This abstract is available from: http://www.journal.chemistrycentral.com/content/3/S1/P64

In chemoinformatics and in the analysis of Quantitative Structure-Activity Relationships (QSAR) experimental data of a molecular property of interest are routinely mathematically related to a set of carefully chosen structure descriptors which represent the molecules under study. Many different mathematical techniques can be used for this purpose. For the sake of simplicity, we focus here on the simplest technique to predict continuous properties, which is linear regression. In a typical model building process the data analyst needs to make several decisions with respect to model complexity and model parameters. Mostly, these decisions are data-driven which makes some form of internal validation necessary to obtain information how the model complexity or model parameters should be set to achieve good predictivity. After the model building process is finished a final validation step with fresh data needs to be carried out to assure that the developed model is truly predictive. The gold standard for doing this is to set aside a portion of the data for this final validation step (the so-called test set validation). There exist several methods to split the original data set into the training and the test set. Common techniques comprise statistical design techniques on the matrix of structure descriptors such as the Kennard-Stone algorithm or the DUPLEX algorithm, using the property vector under study for achieving a uniform distribution in property space, or plain random sampling. All of the aforementioned techniques show some advantages and disadvantages. Here, we highlight the influence of the data splitting algorithm on assessing the predictivity. In recent publications several authors simply focused on the size of the prediction error without accounting for the bias introduced by the splitting algorithm. This is corrected here and it turns out that the commonly accepted best practice is actually not the optimal technique to estimate the true prediction error.

References

1. Baumann K, Albert H, von Korff M: **A systematic evaluation of the benefits and hazards of variable selection in latent variable regression. Part I.** *J Chemom* 2002, **16:**339-350.
2. Snee RD: **Validation of regression models, methods and examples.** *Technometrics* 1977, **19:**415-428.
3. Kennard RW, Stone LA: **Computer aided design of experiments.** *Technometrics* 1969, **11:**137-148.