Data preparation and feature selection for chemical data sets - Building 'general' models

Jörg K. Wegner

Zentrum für Bioinformatik Tübingen (ZBIT), Universität Tübingen, Sand 1, D-72076 Tübingen, Germany

We present some basic principles of data preparation and the actual feature selection topic in 'quantitative structure activity relationship' (QSAR). Especially the high variance for unseen samples (overfitting) will be shortly introduced to grant 'general' models (hypotheses) when applying feature selection.[1,2] We introduce also the relevance of features for similarity analysis and discuss critically the often used 'neighborhood principle' also known as 'structure activity relationship'.[3]

For calculating the descriptors the open source library JOELib was used, which contains all presented descriptor calculation methods.[4]

Examples for predicting the Human Intestinal Absorption (HIA) and octanol/water partition coefficients logP are presented.[1,2]

[1] J.K. Wegner, H. Fröhlich, A. Zell Feature selection for Descriptor based Classification Models. 1. Theory and GA-SEC Algorithm, J. Chem. Inf. Comp. Sci. **2004**; 44; ASAP alert (in print). DOI: 10.1021/ci0342324.

[2] J.K. Wegner, H. Fröhlich, A. Zell A. Feature selection for Descriptor based Classification Models. 2. Human Intestinal Absorption (HIA), J. Chem. Inf. Comp. Sci. **2004**; 44; ASAP alert (in print). DOI: 10.1021/ ci034233w.

[3] N. Nikolova, J. Jaworska. Approaches to Measure Chemical Similarity - a Review, QSAR Comb. Sci. **2003**; 22; 1006-1026; DOI: 10.1002/qsar.200330831.

[4] JOELib, http://joelib.sourceforge.net.

[5] DOI can be directly looked up at http://dx.doi.org