Fechner *et al. Journal of Cheminformatics* 2010, **2**(Suppl 1):P38
http://www.jcheminf.com/content/2/S1/P38

Journal of Cheminformatics

POSTER PRESENTATION **Open Access**

Kernel-based estimation of the applicability domain of QSAR models

Nikolas Fechner[*], Georg Hinselmann, A Jahn, A Zell

From 5th German Conference on Cheminformatics: 23. CIC-Workshop
Goslar, Germany. 8-10 November 2009

Machine learning techniques have become a valuable tool to assess molecular properties without the need of in vitro experiments. Most of these methods do not give any information if a molecule that is predicted can be sufficiently described by the knowledge contained in the model. Thus, the estimation of the reliability of a model-based prediction is an important question in machine learning based QSAR modeling.

One approach to solve this problem is to describe the portion of the chemical space used during the training phase of a model. Any molecule included in the same subspace is then considered as a structure for which the model is regarded as valid. This concept of the description of the subspace in which a model is regarded as reliable is known as the estimation of the *applicability domain* of this model [1].

Most machine learning approaches for QSAR rely on a vectorial representation of the molecules. The applicability domain is expressed as a subspace of the vector space with one dimension for each descriptor used. This concept can be not directly applied to kernel-based techniques like support vector machines. These methods rely on an implicit *feature space* that is only defined by the applied kernel similarity and with unknown dimensions. The applicability domain of a kernel-based model therefore has to be defined by means of the kernel. Consequently, this allows to use structured similarity measures, like the Optimal Assignment Kernel [2] and its extension [3], instead of a numerical encoding. Thus, it is possible to describe the complex chemical structure of many drugs better than it would be using descriptors.

In this work, several approaches to define the applicability domain of a QSAR model by means of a kernel are presented and compared to each other. The approach is to extend the concept of a kernel density estimation to incorporate further information contained in a trained model. This can be achieved by using a weighted average kernel similarity of a predicted molecule to the training data set. The weights can be obtained either by exploiting the knowledge contained in the learned model or by approaches that describe the feature space structure using the kernel.

University of Tübingen, Sand 1, 72076 Tübingen, Germany

Published: 4 May 2010

References

1. Jaworska J, Nikolova-Jeliazkova N, Aldenberg T: *Altern Lab Anim* 2005, **33**:445-459.
2. Fröhlich H, Wegner J, Sieker F, Zell A: *QSAR & Comb Sci* 2006, **25**:317-326.
3. Fechner N, Jahn A, Hinselmann G, Zell A: *J Chem Inf Mod* 2009, **49**:549-560.

doi:10.1186/1758-2946-2-S1-P38
Cite this article as: Fechner *et al.*: Kernel-based estimation of the applicability domain of QSAR models. *Journal of Cheminformatics* 2010 2(Suppl 1):P38.