Chemistry Central Journal

Poster presentation

Open Access

Molecular similarity for machine learning in drug development

M Rupp*, E Proschak and G Schneider

Address: University of Frankfurt, Siesmayerstr. 70, D-60323 Frankfurt am Main, Germany
* Corresponding author

from 3rd German Conference on Chemoinformatics
Goslar, Germany. 11-13 November 2007

Published: 26 March 2008

Chemistry Central Journal 2008, **2**(Suppl 1):P10 doi:10.1186/1752-153X-2-S1-P10

This abstract is available from: http://www.journal.chemistrycentral.com/content/2/S1/P10

In pharmaceutical research and drug development, machine learning methods play an important role in virtual screening and ADME/Tox prediction. For the application of such methods, a formal measure of similarity between molecules is essential. Such a measure, in turn, depends on the underlying molecular representation.

Input samples have traditionally been modeled as vectors. Consequently, molecules are represented to machine learning algorithms in a vectorized form using molecular descriptors. While this approach is straightforward, it has its shortcomings. Amongst others, the interpretation of the learned model can be difficult, e.g. when using fingerprints or hashing.

Structured representations of the input constitute an alternative to vector based representations, a trend in machine learning over the last years. For molecules, there is a rich choice of such representations. Popular examples include the molecular graph, molecular shape and the electrostatic field.

We have developed a molecular similarity measure defined directly on the (annotated) molecular graph, a long-standing established topological model for molecules. It is based on the concepts of optimal atom assignments and iterative graph similarity. In the latter, two atoms are considered similar if their neighbors are similar. This recursive definition leads to a non-linear system of equations. We show how to iteratively solve these equations and give bounds on the computational complexity of the procedure. Advantages of our similarity measure include interpretability (atoms of two molecules are assigned to each other, each pair with a score expressing local similarity; this can be visualized to show similar regions of two molecules and the degree of their similarity) and the possibility to introduce knowledge about the target where available. We retrospectively tested our similarity measure using support vector machines for virtual screening on several pharmaceutical and toxicological datasets, with encouraging results. Prospective studies are under way.