Chemistry Central Journal

Poster presentation

Open Access

A benchmark data set for in silico prediction of ames mutagenicity

K Hansen*, S Mika, T Schroeter, A Sutter, A Ter Laak, T Steger-Hartmann, N Heinrich and K-R Müller

Address: Technical University of Berlin, Franklinstr. 28/29, 10587 Berlin, 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):P31 doi:10.1186/1752-153X-3-S1-P31

This abstract is available from: http://www.journal.chemistrycentral.com/content/3/S1/P31

In silico prediction tools for Ames mutagenicity (*Salmonella typhimurium* reverse mutation assay) represent a cost-effective high throughput approach for the prioritization of compounds before submission to experimental testing. Various modeling approaches have been pursued in this field during the last few years. However, the publicly available data sets used for modeling are mostly very limited in terms of size and chemical coverage. Hence, a reasonable comparison of the different modeling methodologies is so far – as for most QSAR problems – impossible.

In this work we describe a collection of about 6000 non-confidential compounds together with their biological activity in the Ames mutagenicity test. This very large, unique and valuable data set built from public sources is made available in machine-readable form (smiles strings) to be used as a benchmark by other researchers. Based on these data we built three statistical prediction models for Ames mutagenicity based on CORINA and DRAGON descriptors. The methods used are a support vector machine, a random forest and Gaussian processes. All three approaches are evaluated within the same cross-validation setting. To facilitate this valuable benchmark, the exact validation protocol including the exact random splits will be made publicly available. The results show that all three methods yield satisfactory results, reaching sensitivity and specificity values of greater than 70% or 80%, respectively. The application of Gaussian processes, previously not applied to Ames mutagenicity prediction proves slightly superior to the other two methods.