Chemistry Central Journal

Oral presentation

Is learning drugs the same as learning non-drugs?

Robert D Brown* and D Rogers

Address: Accelrys Inc, 10188 Telesis Court, San Diego, CA, USA
* 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):S5 doi:10.1186/1752-153X-2-S1-S5

This abstract is available from: http://www.journal.chemistrycentral.com/content/2/S1/S5

In their recent paper [1] Good and Hermsmeier discuss the effects of test set selection on the evaluation and comparison of SAR methodologies. In particular they examine the impact that analogue effect has on overestimating the predictivness of drug vs non-drug models built by Bayeisan modeling. The analogue effect is a result of most drug and druglike compendia having extensive sets of analogues. When selecting random test sets this tends to ensure that test and training sets both contain members of the same series, which in turn means that the predictivity of a model is greater than it might otherwise be.

Good and Hermsmeier proposed a protocol to evaluate models that reduces this effect, by first organizing a drug database into classes based on the drug ontology classes defined by Schuffenhaur et al. [2]. They then learned models from training sets that excluded a particular class of drug and tested the predictivity on test sets from that class. The authors focused on the ability of various methods to minimize type II errors (false negatives), that is the prediction of drugs as non drugs.

After reproducing their work as far as we are able, we have extended their study to also consider the effects on type I errors (false positives), that is the prediction of a non drug as a drug, which will be an important consideration when one considers the practicality of these methods for selecting sets of samples for synthesis or purchase and screening.

In the reproduction of the original work, we largely concur with the authors that descriptors that encode small and more abstract features of molecules are the most effective at minimizing type II errors. However, minimizing type I errors we found these types of descriptors not to be so effective, and that it was descriptors that encompassed much larger fragments produced the most predictive models. In other words, the descriptors required for learning non-drugs are fundamentally different from those required to learn drugs. We propose therefore that an experiment can be tailored to meet the requirements of precision vs recall by adjusting the environment size that is encoded by the descriptors.

References

1. Good AC, Hermsmeier MA: *J Chem Inf Model* 2007, **47:**110-114.
2. Schuffenhauer A, *et al.*: *J Chem Inf Comput Sci* 2002, **42:**947-955.