Chemistry Central Journal

Poster presentation

Open Access

Distance-dependent: characterizing virtual screening datasets

C Anthes*, SG Rohrer and K Baumann

Address: Institut für Pharmazeutische Chemie, Technische Universität Braunschweig, Beethovenstr. 55, 38106 Braunschweig, 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):P19 doi:10.1186/1752-153X-3-S1-P19

This abstract is available from: http://www.journal.chemistrycentral.com/content/3/S1/P19

Many reports evaluating ligand-based virtual screening methods show that the results are strongly dependent on the composition of the employed benchmark datasets. Recently, it became apparent, that two causes for overoptimistic validation results need to be avoided: artificial enrichment and analogue bias. Artificial enrichment is observed when the decoy set (i. e. the background) differs significantly from the set of actives regarding "simple" molecular properties. Analogue bias describes the fact that in the dataset of actives certain scaffolds are over-represented. Both phenomena render retrieval of actives trivial.

Several techniques were proposed in the literature to cope with these problems. Most of them use the mean of pair wise distances or the mean of pair wise similarity coefficients to characterize dataset diversity [1]. It is obvious that these measures depend on the dataset but also on the employed structure descriptor and the distance/similarity measure.

The goal of this study was to assess whether or not commonly employed measures of diversity reasonably characterize benchmark dataset composition. Therefore, previously published diversity measures were compared to recently introduced spatial statistics-based figures of dataset topology [2]. The relative advantages and disadvantages of the studied figures are contrasted. Interestingly, figures based on more distant neighbours than just the nearest one, performed very well. From a detailed analysis of the findings, a guideline for characterizing ligand-based virtual screening datasets is derived.

References

1. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, Jacoby E, Schuffenhauer A: *J Chem Inf Comput Sci* 2004, **44:**1177-1185.
2. Rohrer SG, Baumann K: *J Chem Inf Model* 2008, **48:**704-718.