Journal of **Chem**informatics

ORAL PRESENTATION   **Open Access**

Molecular bioactivity extrapolation to novel targets by support vector machines

Gerard JP Van Westen[1*], JK Wegner[2], AP IJzerman[1], HWT Van Vlijmen[2], A Bender[1]

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

The early phases of drug discovery use *in silico* models to rationalize structure activity relationships, and to predict the activity of novel compounds. However, the performance of these models is not always acceptable and the reliability of external predictions - both to novel compounds and to related protein targets - is often limited. Proteochemometric modeling [1] adds a target description, based on physicochemical properties of the binding site, to these models.

Our proteochemometric models [2] are based on Scitegic circular fingerprints on the compound side and on a customized protein fingerprint on the target side. This protein fingerprint is based on a selection of physicochemical descriptors obtained from the AAindex database. Through PCA we selected a number of physicochemical properties which are hashed in a fingerprint using the Scitegic hashing algorithm. We compared this fingerprint to a number of protein descriptors previously published, including the Z-scales, the FASGAI and the BLOSUM descriptors. Our fingerprint performs superior to all of these. In addition, we show that proteochemometric models improve external prediction capabilities. In the case of classification this leads to models with a higher specificity when compared to conventional QSAR. In the case of regression our models show an average lower RMSE of 0.12 log units when based on a pIC50 output variable compared to conventional QSAR modeling the same data-set. Furthermore, our models enable target extrapolation. As a result we can predict the activity of known and new compounds on new targets while retaining the same model quality as when performing external validation without target extrapolation.

* Correspondence: gerard@lacdr.leidenuniv.nl
[1]Amsterdam Center for Drug Research, Einsteinweg 55, 2333 CC, Leiden, The Netherlands

Author details
[1]Amsterdam Center for Drug Research, Einsteinweg 55, 2333 CC, Leiden, The Netherlands. [2]Tibotec, Gen De Wittelaan L 11B 3, 2800 Mechelen, Belgium.

References
1. Freyhult E, Prusis P, Lapinsh M, Wikberg JE, Moulton V, Gustafsson MG: Unbiased descriptor and parameter selection confirms the potential of proteochemometric modelling. *BMC Bioinformatics* 2005, **6**:50.
2. Doddareddy MKR, Van Westen GJP, Horst Van der E, Peironcely JE, Corthals F, IJzerman AP, Emmerich M, Jenkins JL, Bender A: Chemogenomics: Looking at Biology through the Lens of Chemistry. *Stat Anal Data Mining* 2009.