Chemistry Central Journal

Chemistry Central

Poster presentation

Virtual screening for PPAR-gamma ligands using the ISOAK molecular graph kernel and gaussian processes

T Schroeter*[1], M Rupp[2], K Hansen[1], K-R Müller[1] and G Schneider[2]

Address: [1]Technische Universität Berlin, Machine Learning Dept., Franklinstr. 28/29, 10587 Berlin, Germany and [2]Johann Wolfgang Goethe-University, Chair for Chem-and Bioinformatics, Siesmayerstr. 70, D-60323 Frankfurt am Main, 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):P15 doi:10.1186/1752-153X-3-S1-P15

This abstract is available from: http://www.journal.chemistrycentral.com/content/3/S1/P15

For a virtual screening study, we introduce a combination of machine learning techniques, employing a graph kernel, Gaussian process regression and clustered cross-validation. The aim was to find ligands of peroxisome-proliferator activated receptor gamma (PPAR-y). The receptors in the PPAR family belong to the steroid-thyroid-retinoid superfamily of nuclear receptors and act as transcription factors. They play a role in the regulation of lipid and glucose metabolism in vertebrates and are linked to various human processes and diseases [1]. For this study, we used a dataset of 176 PPAR-y agonists published by Ruecker et al [2].

Gaussian process (GP) models can provide a confidence estimate for each individual prediction, thereby allowing to assess which compounds are inside of the model's domain of applicability. This feature is useful in virtual screening, where a large fraction of the tested compounds may be outside of the model's domain of applicability. In cheminformatics, GPs have been applied to different classification and regression tasks using either radial basis function or rational quadratic kernels based on vectorial descriptors [4,5]. We used a graph kernel based on iterative similarity and optimal assignments (ISOAK, [3]) for non-linear Bayesian regression with Gaussian process priors (GP regression, [4]). A number of kernel-based learning algorithms (including GPs) are capable of multiple kernel learning [5], which allows combining heterogeneous information by using multiple kernels at the same time. In this work, we combined rational quadratic kernels for vectorial molecular descriptors (MOE2D, CATS2D and Ghose-Crippen fragment descriptors) with the ISOAK graph kernel.

We evaluated our methodology in different ranking and regression settings. Ranking performance was assessed using the number of false positives within the top k predicted compounds. Predicted compounds were ranked based on both predicted binding affinity and the confidence in each prediction. In the regression setting, we employed standard loss functions like mean absolute error (MEA) and root mean squared error. The established linear ridge regression (LRR) and support vector regression (SVR) algorithms served as baseline methods. In addition to standard test/training splits and cross-validation, we used a clustered cross-validation strategy where clusters of compounds are left out when constructing training sets. This results in less optimistic results, but has the advantage of favouring more robust and potentially extrapolation-capable algorithms than standard training/test splits and normal cross-validation. In the regression setting, both GP and SVR models performed well, yielding MAEs as low as 0.66 +- 0.08 log units (clustered CV) and 0.51 +- 0.3 log units (normal CV). In the ranking setting, GPs slightly outperform SVR (0.21 +- 0.09 log units vs. 0.3 +- 0.08 log units).

In conclusion, Gaussian process regression using simultaneously – via multiple kernel learning – the ISOAK molecular graph kernel and the rational quadratic kernel (with standard molecular descriptors) performs excellent in retrospective evaluation. A prospective evaluation study is currently in progress.

References

1. Henke B: *Progr Med Chem* 2004:1-53.
2. Ruecker C, Scarsi M, Meringer M: *Bioorg Med Chem* 2006:5178-5195. Rupp M, Proschak E, Schneider G, *J Chem Inform Model*, 2007, 2280–2286.
3. Schwaighofer A, Schroeter T, Mika S, Laub J, Laak A, Sülzle D, Ganzer U, Heinrich N, Müller K-R: *J Chem Inform Model* 2007:407-424.
4. Obrezanova O, Csanyi G, Gola J, Segall M: *J Chem Inf Model* 2007:1847-1857.
5. Sonnenburg S, Rätsch G, Schäfer C, Schölkopf B: *J Machine Learning Research* 2006:1531-1565.