

ORAL PRESENTATION

Open Access

Graph kernels for chemoinformatics—a critical discussion

Matthias Rupp

From 6th German Conference on Chemoinformatics, GCC 2010
Goslar, Germany. 7-9 November 2010

We analyze the use, advantages, and drawbacks of graph kernels in chemoinformatics, including a comparison of kernel-based approaches with other methodology, as well as examples of applications.

Kernel-based machine learning [1], now widely applied in chemoinformatics, delivers state-of-the-art performance [2] in tasks like classification and regression. Molecular graph kernels [3] are a recent development where kernels are defined directly on the molecular structure graph. This allows the adaptation of methods from graph theory to structure graphs and their direct use with kernel learning algorithms. The main advantage of kernel learning, the so-called “kernel trick”, allows for a systematic, computationally feasible, and often globally optimal search for non-linear patterns, as well as the direct use of non-numerical inputs such as strings and graphs. A drawback is that solutions are expressed indirectly in terms of similarity to training samples, and run-times that are typically quadratic or cubic in the number of training samples.

Graph kernels [3] are positive semidefinite functions defined directly on graphs. The most important types are based on random walks, subgraph patterns, optimal assignments, and graphlets. Molecular structure graphs have strong properties that can be exploited [4], e.g., they are undirected, have no self-loops and no multiple edges, are connected (except for salts), annotated, often planar in the graph-theoretic sense, and their vertex degree is bounded by a small constant. In many applications, they are small. Many graph kernels are general-purpose, some are suitable for structure graphs, and a few have been explicitly designed for them.

We present three exemplary applications of the iterative similarity optimal assignment kernel [5], which was designed for the comparison of small structure graphs: The discovery of novel agonists of the peroxisome proliferator-activated receptor γ [6] (ligand-based virtual screening), the estimation of acid dissociation constants [7] (quantitative structure-property relationships), and molecular de novo design [8].

Published: 19 April 2011

References

1. Hofmann T, Schölkopf B, Smola A: Kernel methods in machine learning. *Ann Stat* 2008, **36**(6):1171-1220.
2. Invauciu O: Applications of Support Vector Machines in Chemistry. In *Reviews in Computational Chemistry* 23. Weinheim: Wiley-VCH;Lipkowitz K, Cundari T 2007:291-400.
3. Rupp M, Schneider G: Graph kernels for molecular similarity. *Mol Inf* 2010, **29**(4):266-273.
4. Rupp M: Kernel Methods for Virtual Screening. *PhD thesis* 2009.
5. Rupp M, Proschak E, Schneider G: Kernel approach to molecular similarity based on iterative graph similarity. *J Chem Inf Model* 2007, **47**(6):2280-2286.
6. Rupp M, Schroeter T, Steri R, Zettl H, Proschak E, Hansen K, Rau O, Schwarz O, Müller-Kuhrt L, Schubert-Zsilavecz M, Müller K-R, Schneider G: From machine learning to natural product derivatives selectively activating transcription factor PPAR γ . *ChemMedChem* 2010, **5**(2):191-194.
7. Rupp M, Körner R, Tetko I: 2010, submitted.
8. Hartenfeller M, Schneider G: De novo drug design. *Meth Mol Biol* 2011, **672**:299-323.

doi:10.1186/1758-2946-3-S1-O8

Cite this article as: Rupp: Graph kernels for chemoinformatics—a critical discussion. *Journal of Cheminformatics* 2011 **3**(Suppl 1):O8.

Correspondence: mrupp@mrupp.info
Helmholtz Zentrum München, Deutsches Forschungszentrum für Umwelt & Gesundheit, Ingolstädter Landstr. 1, 85764 Neuherberg, Germany