Journal of Cheminformatics

POSTER PRESENTATION Open Access

# Maximum-score diversity selection for early drug discovery

Thorsten Meinl*, C Ostermann, O Nimz, A Zaliani, MR Berthold

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

Diversity selection is a common task in early drug discovery, be it for removing redundant molecules prior to HTS or reducing the number of molecules to synthesize from scratch. One drawback of the current approach, especially with regard to HTS, is, however, that only the structural diversity is taken into account. The fact that a molecule may be highly active or completely inactive is usually ignored. This is especially remarkable, as quite a lot of research is involved in improving virtual screening methods in order to forecast activity. We therefore present a modified version of diversity selection – which we termed Maximum-Score Diversity Selection – which additionally takes the predicted activities of the molecules into account. Not very surprisingly both objectives – maximizing activity whilst also maximizing diversity in the selected subset – conflict. As a result, we end up with a multiobjective optimization problem. We will show, that the task of diversity selection is quite complicated (it is NP-complete) and therefore heuristic approaches are needed for typical dataset sizes.

A common and popular approach is using multiobjective genetic algorithms, such as NSGA-II [1], for optimizing both objectives for the selected subsets. However, we will show that usual implementations suffer from severe limitations that prevent them from finding quite a lot of possible interesting solutions. Therefore, we evaluated two other heuristic for maximum-score diversity selection. One is special heuristic (called BB2) that was motivated by the mentioned proof of NP-completeness [2]. The other is a novel heuristics called Score Erosion which was specifically developed for our actual problem. Among all three heuristics, Score Erosion is by far the fastest one while finding solutions of equal quality compared to the genetic algorithm and BB2. This will be shown on several real world datasets, both public and internal ones.

All experiments were carried out using the data analysis platform KNIME [3] therefore we will also show some example how maximum-score diversity selection can be performed inside workflow-based environments.

University of Konstanz, 78457 Konstanz, Germany

Published: 4 May 2010

**References**

1. Deb K, Pratap A, Agarwal S, Meyarivan T: **A fast and elitist multiobjective genetic algorithm: NSGA-II.** *IEEE Transactions on Evolutionary Computation* 2002, **6**:182-197.
2. Erkut E: **The discrete p-dispersion problem.** *European Journal of Operational Research* 1990, **46(1)**:48-60.
3. [http://www.knime.org/].

doi:10.1186/1758-2946-2-S1-P33
**Cite this article as:** Meinl *et al.*: **Maximum-score diversity selection for early drug discovery.** *Journal of Cheminformatics* 2010 **2**(Suppl 1):P33.