Chemistry Central Journal

Poster presentation

Open Access

Complexity effects in fingerprint similarity searching

Y Wang, H Geppert and J Bajorath*

Address: Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universtität Bonn, Dahlmannstr. 2, D-53113 Bonn, 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):P5 doi:10.1186/1752-153X-3-S1-P5

This abstract is available from: http://www.journal.chemistrycentral.com/content/3/S1/P5

Similarity searching using fingerprint representations of molecules is widely applied for mining of chemical databases [1]. Known active compounds are used as templates to search for novel hits using similarity measures for quantitative bit string comparison. A variety of similarity metrics are being used for this purpose including the popular Tanimoto coefficient [1] and the Tversky coefficients [2].

Differences in molecular complexity and size are known to bias the evaluation of fingerprint similarity [3]. Complex molecules tend to produce fingerprints with higher bit density than simpler ones, which often leads to artificially high similarity values in search calculations. For example, we have thoroughly analyzed similarity value distributions and demonstrated that apparent asymmetry in Tversky similarity search calculations is a direct consequence of differences in fingerprint bit densities [4].

There are in principle two approaches to balance complexity effects; either by designing fingerprints that have constant bit density, regardless of the nature of test molecules, or, alternatively, by introducing similarity metrics that equally weight bit positions that are set on or off. We have shown that a size-independent fingerprint with constant bit density does not produce asymmetrical search results [4]. In addition, a novel similarity metric has been developed, which not only balances complexity effects, but also results in further improved search performance compared to conventional calculations on Tanimoto similarity [5]. However, highly complex molecules are generally much less suitable as reference compounds for fingerprint searching than active compounds having complexity comparable to the screening database [5]. Random deletion of bits that are set on in complex templates has been shown to increase compound recall, despite the associated loss in chemical information content [6]. Taking relative chemical complexity of reference and database compounds into account makes it possible to increase the success rates of fingerprint similarity searching.

References

1. Willett P, *et al.*: *J Chem Inf Comput Sci* 1998, **38(6):**983-96.
2. Chen X, Brown F: *Chem Med Chem* 2007, **2(2):**180-2.
3. Flower D: *J Chem Comput Sci* 1998, **38(3):**379-86.
4. Wang Y, *et al.*: *Chem Med Chem* 2007, **2(7):**1037-42.
5. Wang Y, Bajorath J: *J Chem Inf Model* 2008, **48(1):**75-84.
6. Wang Y, *et al.*: *Chem Biol Drug Design* 2008, **71(6):**511-7.