Chemistry Central Journal

ChemistryCentral

Poster presentation

Open Access

Distance phenomena in high-dimensional chemical descriptor spaces: consequences for similarity-based approaches

M Rupp* and G Schneider

Address: University of Frankfurt, Siesmayerstr. 70, D-63791 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):P28 doi:10.1186/1752-153X-3-S1-P28

This abstract is available from: http://www.journal.chemistrycentral.com/content/3/S1/P28

Measuring the (dis)similarity of molecules is, besides descriptor selection, an important factor for many cheminformatics applications like compound ranking, clustering, and, property prediction. In this work, we focus on real-valued vector spaces (as opposed to the binary spaces of, e.g., fingerprints). We demonstrate the severe influence the choice of (dis)similarity measure can have on the results of cheminformatics applications, and provide recommendations for such choices.

We briefly review the mathematical concepts [1] used to measure (dis)similarity in vector spaces, namely norms, metrics, inner products and similarity coefficients, and the relationships between them, employing commonly used [2][3] (dis)similarity measures in cheminformatics as examples.

Then, we present several phenomena (empty space phenomenon, sphere volume related phenomena, distance concentration [4][5][6]) in high-dimensional descriptor spaces which are not encountered in two and three dimensions. These phenomena are theoretically characterized and illustrated with both artificial and real (bioactivity) data examples.

References

1. Meyer C: *Matrix Analysis and Applied Linear Algebra, SIAM, Philadelphia* 2001.
2. Leach A, Gillet V: *An Introduction to Chemoinformatics Springer Netherlands*; 2003.
3. Willett P: *J Chem Inf Comput Sci* 1998, **38:**983-996.
4. Aggarwal C, Hinneburg A, Keim D: *ICDT 2001 Proceedings, 2001, LNCS* 1973:420-434.
5. Beyer K, Goldstein J, Ramakrishnan R, Shaft U: *ICDT 1999 Proceedings, LNCS 1540* 1999:217-235.
6. Francois D, Wertz V, Verleysen M: *IEEE Trans Knowl Data Eng* 2007, **19:**873-886.