Large-scale comparison of similarity metrics for molecular and interaction fingerprints

Dávid Bajusz, Anita Rácz, Károly Héberger
Research Centre for Natural Sciences, Hungarian Academy of Sciences, Budapest, Hungary

Molecular fingerprints are ubiquitously applied to represent molecular structure in a wide range of applications in cheminformatics, computational drug discovery and related fields [1]. Their greatest advantages are their compactness and machine-readability: the latter enables the fast comparison of molecular structure and the quantification of their similarity. Similarly, interaction fingerprints encode information about protein-ligand complexes (highly relevant in drug discovery) in an equally compact manner [2].

However, there are a great number of methods for calculating the similarity of two such binary data structures and – despite the general preference of the cheminformatics community towards the Tanimoto coefficient – the choice of similarity metric is not trivial.

Recently, we have shown with robust statistical methods on a large dataset that the Tanimoto coefficient is a justified choice from a small pool of commonly known and easily available similarity metrics – although other, equally consistent metrics could be identified as well [3]. In 2012, Todeschini and coworkers have compared a greater number of similarity metrics using a different methodology on simulated and real virtual screening datasets [4].

Currently, we are extending our methodology to a larger pool of similarity metrics [1,4] and molecular, as well as interaction fingerprints, implemented in various cheminformatics and modeling packages. Similarity metrics will be compared and ranked based on their consistency with a suitable reference method (data fusion). We also plan on releasing an open-source Python module implementing the studied similarity metrics, which will be readily applicable with the Cinfony toolkit, the open-source aggregator of cheminformatics software [5].

Acknowledgement

The authors thank the support of the National Research, Development and Innovation Office of Hungary (OTKA contracts K 119269 and KH-17 125608).

References

[1] Bajusz, D.; Rácz, A.; Héberger, K. Chemical Data Formats, Fingerprints, and Other Molecular Descriptions for Database Analysis and Searching. In: Comprehensive Medicinal Chemistry III; Chackalamannil, S.; Rotella, D.P.; Ward, S.E., Eds.; Elsevier: Oxford, 2017; pp. 329–378.

[2] Vass, M.; Kooistra, A.J.; Ritschel, T.; Leurs, R.; de Esch, I.J.; de Graaf, C. Molecular Interaction Fingerprint Approaches for GPCR Drug Discovery. Curr. Opin. Pharmacol., 2016, 30, 59–68.

[3] Bajusz, D.; Rácz, A.; Héberger, K. Why Is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations? J. Cheminform., 2015, 7, 20.

[4] Todeschini, R.; Consonni, V.; Xiang, H.; Holliday, J.; Buscema, M.; Willett, P. Similarity Coefficients for Binary Chemoinformatics Data: Overview and Extended Comparison Using Simulated and Real Data Sets. J. Chem. Inf. Model., 2012, 52, 2884–2901.

[5] O'Boyle, N.M.; Hutchison, G.R. Cinfony – Combining Open Source Cheminformatics Toolkits behind a Common Interface. Chem. Cent. J., 2008, 2, 24.