Evaluation of Extended Similarity Indices for Molecule Similarity and Chemical Similarity Networks

Timothy Dunn[1], Chenglong Li[2], Gustavo Seabra[2] and Ramon A Miranda Quintana[1]

[1]Department of Chemistry, University of Florida, USA, Email: tdunn1@ufl.edu
[2]College of Pharmacy, University of Florida, USA

Cheminformatics is an ever-growing field of data science. It involves the utilization of large volumes of data, experimental or theoretical, to analyze and describe complex relations in chemistry with the aid of computers. One of its major applications is in the discovery of new small molecules for drug design. However, determination of the similarity indices between large sets of molecules is a resource intensive process due to the need to compare each pair of molecules.

Recent progress has been made in extended similarity indices that allow for the simultaneous comparison of molecules in a dataset. This reduces the scale of the problem from $O(N^2)$ to $O(N)$ which drastically speeds up computation times for large data sets. Previous work defined these extended similarity indices and showed their improved performance with large sets of molecules over traditional similarity indices as well as demonstrated their potential application in clustering.

This presentation aims to further expand upon these initial findings. Chemical similarity and diversity as defined by these extended similarity indices is evaluated with different fingerprint types and coincidence thresholds. The effect of randomized datasets on these extended similarity indices is also explored. The application of these indices to chemical similarity networks is also discussed.