Abstract

Methods for measuring the similarity of mass spectra and the structures of small molecules are crucial for advancements in medicinal chemistry, pharmacology, and metabolomics. One commonly used method for comparing the mass spectra of molecules is cosine similarity. This measures the similarity between two non-zero vectors by calculating the cosine of the angle between them. Comparing the mass spectra of molecules enables searching in molecular databases, clustering of spectra, and exploration of spectral libraries. Structural similarity is measured based on various molecular fingerprints, such as Daylight, RDKit, Atom-Pair, Topological Torsion, Extended-Connectivity fingerprints, and others. These fingerprints are compared using similarity coefficients. The methods for comparing structures and mass spectra of molecules mentioned can be applied using bioinformatic libraries such as RDKit and CDK for generating and analyzing structural fingerprints, and the MatchMS library for comparing mass spectra. The work provides a theoretical overview of molecular descriptors, including various types of molecular fingerprints and techniques for measuring structural similarity, as well as the principles of mass spectrometry and approaches to comparing mass spectra. The practical part of the work focuses on analyzing structural and spectral data obtained from the MoNA database (MassBank of North America) using RDKit, CDK, and MatchMS