Abstract

Shotgun proteomics is a widely known technique for identification of protein and peptide sequences from an "in vitro" sample. A tandem mass spectrometer generates tens of thousands of mass spectra which must be annotated with peptide sequences. For this purpose, the similarity search in a database of theoretical spectra generated from a database of known protein sequences can be utilized. Since the sizes of databases grow rapidly in recent years, there is a demand for utilization of various database indexing techniques. We investigate the capabilities of (non)metric access methods as the database indexing techniques for fast and approximate similarity retrieval in mass spectra databases. We show that the method for peptide sequences identification is more than 100x faster than a sequential scan over the entire database while more than 90% of spectra are correctly annotated with peptide sequences. Since the method is currently suitable for small mixtures of proteins, we also utilize a precursor mass filter as the database indexing technique for complex mixtures of proteins. The precursor mass filter followed by ranking of spectra by a modification of the parametrized Hausdorff distance outperforms state-of-the-art tools in the number of identified peptide sequences and the speed of search. The proposed methods are implemented in the peptide identification engine SimTandem which can be used for a batch analysis in the framework TOPP based on OpenMS.