Algorithms and statistical models for predicting molecules fragmentation scheme during collision-induced dissociation

Mass spectrometry is an experimental method used in chemistry for measurement of weight of molecules. It is a very useful tool with a very broad application in many fields from agriculture, to pharmaceutical, medical and forensic analysis. It helps in the identification and characterization of unknown compounds (either proteins or the smaller ones like metabolites), development and design of new compounds such as pharmaceutical drugs. Tandem mass spectrometry is one of the typical mass spectrometry experiment setups, which consists of two mass spectrometers connected one by one. In the first mass spectrometer, ionized molecules are separated based on their mass to charge ratio (m/z) and then detected. In the second mass spectrometer ions with selected m/z values are fragmented and their fragments are detected. A popular technique of fragmentation is the collision-induced dissociation (CID). It consists of accelerating molecules to higher kinetic energy and then colliding them with low-energetic, neutral atoms of noble gases like helium. The study of fragmentation schemes is an interesting research topic because it helps in better understanding of structures of molecules. While CID fragmentation is well characterized for proteins, it is still a challenging problem for metabolites and other small molecules. This is caused by their complicated, non-linear structure, as opposed to the regular, linear structure of proteins. One of the important problems of mass spectrometry, which still requires further investigation, is the automatic elucidation of the fragmentation mechanism for a given molecule. A solution to this problem requires an application of sophisticated algorithmic and statistical approach due to its aforementioned complexity.

The main objective of this project is a computational software tool which predicts the fragmentation scheme for a given precursor chemical structure and its spectrum. On the one hand, the program should propose the most feasible fragmentation scheme, while on the other hand, it should explain the given spectrum as well as possible. The tool is going to be able to be used in a daily laboratory routine thanks to easily understandable output and simple interface.

We propose a combinatorial solution to this problem, which consists of in-silico simulating feasible fragmentation reactions and choosing those which optimally explain the mass spectrum. The whole space of all possible reactions is too big to be simulated by a typical computer. We use algorithmic techniques such as priority queues and heuristic score functions to simulate the most feasible reactions. The algorithms are inspired by those used in artificial intelligence. The optimal reaction mechanism is then chosen from the simulated reactions to explain as much of the spectrum as possible in the simplest possible way.

Our aim is to analyze a fairly wide range of possible fragmentations using a few simple rules. For this reason, we describe the reaction as a sequence of simple reaction patterns: bond splits, rearrangement of multiple bonds, and joins of fragments. Even though joins do not occur in actual fragmentation processes they are used to simulate complex reactions which involve several atoms from different parts of the molecule, such as McLafferty rearrangement.

Initial implementation of a prototype of the software tool shows that our approach gives more feasible fragmentation compared to the approach known from other popular tools. Including a set of elementary reactions allows for better spectrum explanation and better reflects real fragmentation processes. However, there is still a broad space for developing algorithms and tools which propose good spectrum explanation with reasonable computational complexity.