Abstract

This PhD project is on the algorithmic aspects of synthesis planning and mass spectrometry; two separate chemical problems concerning the understanding of molecules and molecular interactions. From a modeling perspective, the goal of synthesis planning is to assemble smaller molecules into larger ones, while mass spectrometry is concerned with how molecules break into smaller fragments. For this reason the thesis is divided into two parts: Part I explores synthesis planning, while Part II is on mass spectrometry.

Part I: In synthesis planning, the goal is to synthesize a target molecule from available starting materials, possibly optimizing costs such as price or environmental impact of the process. Current algorithmic approaches to synthesis planning are usually based on selecting a bond set and finding a single good plan among those induced by it. We demonstrate that synthesis planning can be phrased as a combinatorial optimization problem on hypergraphs. For this, individual synthesis plans are modeled as directed hyperpaths embedded in a hypergraph of reactions (HoR) representing the chemistry of interest. As a consequence, a polynomial time algorithm to find the K shortest hyperpaths can be used to compute the K best synthesis plans for a given target molecule. Having K good plans to choose from has several benefits: It makes the synthesis planning process much more robust when in later stages adding further chemical details, it allows one to combine several notions of cost, and it provides a way to deal with imprecise yield estimates.

For computation of optimal synthesis plans, classical quality measures are discussed: Total Weight of Starting Materials has previously been used for ranking synthesis plans in different ways. The connection between the different definitions is shown, and the quality measure is applied to our method. The quality measure External Path Length, an expression of convergency, has a certain peculiarity that makes its use for ranking synthesis plans questionable.

An empirical study of our method illustrates the limitations of what a chemist can expect is feasible to compute, and how ranking of plans may vary according to different values of yield. The latter demonstrates the practical value of our method for cases where yield estimates are imprecise or unknown. To illustrate the realism of the approach, synthesis plans from our abstraction level are compared with detailed chemical synthesis plans from the literature. For this, a synthesis plan for Wieland-Miescher ketone and a synthesis plan for lysergic acid are used.

A bond set gives rise to a HoR in a natural way. However, our modeling is not restricted to bond set based approaches—any set of known reactions and starting materials can be used to define a HoR. We do, however, describe an algorithm for constructing a HoR efficiently, for cases where the HoR is not known in advance. Furthermore, equivalence of our structural definition of a hyperpath and two definitions from the hypergraph literature is shown.

Part II: Mass spectrometry is an analytic technique for characterizing molecules and molecular mixtures, by gaining knowledge of their structure and composition from the way they fragment. In a mass spectrometer, molecules or molecular mixtures are ionized and fragmented, and the abundances of the different fragment masses are measured, resulting in so-called mass spectra.

We suggest a new road map improving the current state-of-the art in computational methods for mass spectrometry, mainly in prediction of mass spectra. Our main focus is on increasing the chemical realism of the modeling of the fragmentation process. Two core ingredients of our proposal are i) describing the individual fragmentation reactions via graph transformation rules and ii) expressing the dynamics of the system via reaction rates and quasi-equilibrium theory. We use graph transformation rules both for specifying the possible core fragmentation reactions, and for characterizing the reaction sites when learning values for the rates. We believe that this model describes chemical mechanisms more accurately than previous ones, and that this can lead to both better spectrum prediction and more explanatory power. Our modeling of system dynamics also allows better separation of instrument dependent and instrument independent parameters of the model.