Analysis of Generative Chemistries

Dissertation, 2015

Jakob Lykke Andersen

For the modelling of chemistry we use undirected, labelled graphs as explicit models of molecules and graph transformation rules for modelling generalised chemical reactions. This is used to define artificial chemistries on the level of individual bonds and atoms, where formal graph grammars implicitly represent large spaces of chemical compounds. We use a graph rewriting formalism, rooted in category theory, called the Double Pushout approach, which directly expresses the transition state of chemical reactions. Using concurrency theory for transformation rules, we define algorithms for the composition of rewrite rules in a chemically intuitive manner that enable automatic abstraction of the level of detail in chemical pathways. Based on this rule composition we define an algorithmic framework for generation of vast reaction networks for specific spaces of a given chemistry, while still maintaining the level of detail of the model down to the atomic level. The framework also allows for computation with graphs and graph grammars, which is utilised to model non-trivial chemical systems. The graph generation relies on graph isomorphism testing, and we review the general individualisation-refinement paradigm used in the state-of-the-art algorithms for graph canonicalisation, isomorphism testing, and automorphism discovery.

We present a model for chemical pathways based on a generalisation of network flows from ordinary directed graphs to directed hypergraphs. The model allows for reasoning about the flow of individual molecules in general pathways, and the introduction of chemically motivated routing constraints. It further provides the foundation for defining specialised pathway motifs, which is illustrated by defining necessary topological constraints for both catalytic and autocatalytic pathways. We also prove that central types of pathway questions are NP-complete, even for restricted classes of reaction networks. The complete pathway model, including constraints for catalytic and autocatalytic pathways, is implemented using integer linear programming. This implementation is used in a tree search method to enumerate both optimal and near-optimal pathway solutions.

The formal methods are applied to multiple chemical systems: the enzyme catalysed beta-lactamase reaction, variations of the glycolysis pathway, and the formose process. In each of these systems we use rule composition to abstract pathways and calculate traces for isotope labelled carbon atoms. The pathway model is used to automatically enumerate alternative non-oxidative glycolysis pathways, and enumerate thousands of candidates for autocatalytic pathways in the formose process.