LOTUS structural expansion

Pascal Amrein

Master thesis in Bioinformatics and Computational Biology

Exploring the chemical space of natural products (NPs) is crucial for the discovery of new bioactive compounds. This study expands the LOTUS dataset, a comprehensive database of over 750,000 structural organism pairs, by using Pickaxe software to predict novel metabolic reactions and compounds. After selecting unique starting compounds and removing stereochemistry, approximately 140,000 molecules from LOTUS were available for prediction with Pickaxe, resulting in the generation of 3.2 million new potential molecules and 3.4 million new reactions based on 250 chemical rules. For the input of molecules, SMILES (Simplified Molecular Input Line Entry System) was used to represent chemical structures and SMARTS (SMiles ARbitrary Target Specification) was used to encode the chemical transformation rules. The methodology focused on the use of in silico techniques to generate these predictions, which were stored and analysed in a Mongo database. A notable result was the frequent prediction of bromine, with bromine being the most common element among the newly predicted molecules. This frequent occurrence can be attributed to the versatility of bromine in the formation of stable compounds and its known role in biological systems, e.g. as a cofactor in enzymatic reactions. Principal component analysis (PCA) showed a broad distribution of the predicted compounds around the starting compounds, indicating the plausibility of the reactions used. These results emphasise the potential of computational tools in expanding the chemical space of natural products and represent a valuable resource for future drug discovery. As a next step, the generation of mass spectrometry (MS) spectra for the predicted compounds and their comparison with experimentally generated spectra by untargeted metabolomics could help in the discovery of new molecules.

Supervisor: Pierre-Marie Allard