Predicting Mass Spectra with Neural Networks

Cameron Mccosh - 2242033m

December 20, 2019

1 Status report

1.1 Proposal

1.1.1 Motivation

Mass spectrometry (MS) is widely used in microbiology to break down samples and identify atoms and molecules by their mass or characteristic fragmentation pattern. In metabolomics, which is the study of molecules that are the products and intermediates of metabolism, MS is used to break down metabolites and produce spectra which can be used to identify molecules. Metabolites have complex structures which means that they are inherently difficult to identify and many are still unknown. MS is expensive and is a high-throughput method which produces a large amount of data which requires automated methods of searching to match spectra to molecules. Database searching is the easiest method of identification but requires that a molecule's spectra has previously been identified and recorded, which is not the case for many metabolites.

1.1.2 Aims

This project aims to predict mass spectra from just the molecule's fingerprint in SMILES format in order to augment spectra databases with synthetic spectra obtained by a neural network. This would allow researchers to use database searching to match MS data with previously unidentified metabolites.

1.2 Progress

- Conducted background researched of topic to gain an understanding of the field
- Decided on python as language of choice and Keras as the deep learning library
- Received data to use and implemented functions to read and validate data
- Set up CI environment to monitor builds

1.3 Problems and risks

1.3.1 Problems

- Unfamiliar with the field and took a while to understand the problem and how to get started

- Not much experience with deep learning and neural networks, Keras is a fairly easy library to get started with but still took some time to get to grips with it
- Time constraints, overextended myself this semester

1.3.2 Risks

- Data may not be enough to effectively train a neural network - will start small by predicting single spectra and work up
- Evaluating results may be difficult - I will plot predictions against real spectra and determine a metric to use to measure performance

1.4 Plan

- Week 1: have single spectrum prediction working
- Weeks 2-5: increase complexity of neural network to include multiple spectra and intensities
- Weeks 6-7: run evaluations
- Weeks 8-10: write up dissertation and prepare presentation