NMR Spectra Structure Relationship

Scott Thiel
thiels2@rpi.edu

September 2018

Overview

Nuclear Magnetic Resonance (NMR) spectroscopy is an important spectroscopic technique for observing local magnetic fields of atomic nuclei. This yields valuable information about the environment which assists in the identification and determination of the molecular species in question. The spectra of small molecules are often easily predictable, and the presence of most functional groups is obvious. Some molecules produce spectra that may be very busy however, which makes identification an arduous task. In the case of novel molecules that have yet to be identified and their spectrum published, it may not be possible to search for them in a database or in some cases feasible to generate a spectrum. This project proposes an alternative method to connecting molecules to their NMR spectra.

Description

The NMR spectrum a molecule generates is a result of specific local nuclear magnetic properties of that molecule which essentially means that to some degree there exists a definitive relationship between the two. It is then proposed that a sufficiently sized neural network, given enough data, should eventually be able to model this relationship. Once modeled, this relationship between the NMR spectrum and the structure of a molecule can be used to either generate a spectrum given a molecule, or perhaps generate a molecule or set of molecules given a spectrum.

Model

The first aspect to be implemented would be initial processing to get NMR spectral data and molecular structure in a form suitable for a neural network. On the spectra side, this would involve parsing xml and placing peak intensities into an array whose indices are scaled chemical shifts. As for the structure, the three methods that would likely prove most useful are chemical smiles, topological fingerprints, and convolution graphs. The former two are to be implemented first as they are less complex and more easily implemented than the latter. One the available data can be effectively processed, construction of the neural network will follow. The first implementation of a network would take a bit array representation of the molecule as input, be that a smile, fingerprint, or graph, and output a vector of intensities indexed by chemical shift. Output would be compared to the molecule's experimental spectrum likely using either the Euclidian norm or the cosine similarity. As before, the trained model would take molecular structure as input and yield an artificial NMR spectrum that would ideally resemble the experimental one.

More info can be found at the project repsoitory: `https://github.com/pdmlc/nssr`