Poster K-20

From Spectrum to Structure using Machine Learning

Authors:
Caroline Farrelly (*University of Manchester*)
Prof. Douglas Kell (*University of Manchester*)
Dr. Joshua Knowles (*University of Manchester*)

Short Abstract: Current trends in the use of high throughput techniques have created demand for rapid analytical techniques. In the areas of combinatorial chemistry and metabolomics alike, huge datasets are generated, often relating to novel compounds. This project addresses structure elucidation of carbon NMR spectra for small organic molecules (>500 molecular weight).

Long Abstract:

Current trends in the use of high throughput techniques have created demand for rapid analytical techniques. In the areas of combinatorial chemistry and metabolomics alike, huge datasets are generated, often relating to novel compounds. This project addresses structure elucidation of carbon NMR spectra for small organic molecules (>500 molecular weight).

NMR is a powerful analytical tool, able to generate spectra from tiny sample sizes; in as little as a few seconds for protons. However, spectral analysis is time consuming and traditionally involved high levels of human expertise. This has fuelled the development of automated techniques for structure elucidation from NMR spectra.
Previous efforts have focused on empirical methods or expert systems, with some work in the semi-empirical area. This project offers a fresh perspective by using Ant Colony Optimisation methods to learn associations between spectral peaks and substructural features.

When a molecule is placed in an external magnetic field, each nucleus will give rise to a different spectral peak according to its electronic environment. The positioning of peaks on the spectrum depends on the external field strength, combined with the total secondary magnetic field, so they are best described as chemical shifts rather than absolute values. (Hornak 1997). Certain shift ranges are associated with functional groups, but these are distorted by adjacent environments.

Automated spectrum prediction is an established field and has been developed as an aid in structure elucidation. By predicting the spectrum for an expected molecule a chemist can quickly establish any similarity between the two.

The inverse problem of automated structure elucidation is much more computationally challenging. The secondary magnetic fields mean that chemical shifts are dependant upon a molecules total electronic configuration. Subsequently conventional algorithms are not suited to the levels of pattern recognition required for spectral analysis, so artificial intelligence methods are relied upon.

Although expert systems and neural networks have been applied with some success to this problem, swarm intelligence methods remain largely unexplored (see Gasteiger, 2003). The main benefit of this type of solution would be a matrix system that could be reused independent of the original dataset used to develop it.

Starting off with a dataset of ~3000 fully assigned C13 NMR Spectra, each molecule has been broken down into its substructural components in order that a substructural matrix may be constructed revealing any correlations between functional groups and chemical shifts. By varying substructure size and looking at the strength of electronic environment, patterns can begin to be identified.

The machine learning phase of this project uses an ACO program to use "building blocks" from the substructural matrix to generate isomers, evaluating each one of these against a fitness function. During the search, the probabilities of certain substructures are reinforced by building a pheromone matrix to reflect which substructures appear more often in the best isomers. This pheromone matrix is used to attract more ants to substructures with higher pheromone levels.

Ant Colony Optimisation (ACO) algorithms are the result of modeling the behaviour of ants foraging for a food source and finding the shortest path to return to their colony. Algorithms use artificial ants, initially taking random search paths, depositing pheromone trails as they travel. Those ants who have taken the shorter path will return to the nest more quickly, so a stronger scent pheromone will remain on this path. Hence, more ants will be attracted to take this same path. Some algorithms allow for pheromone evaporation, making the system more flexible in case of the emergence of new paths or closure of existing paths. Ant “memory” has been incorporated into some models. In this case the ants remember the location of their nest and so have a better sense of direction, so can adjust their path accordingly. (Bonabeau et al. 1999)

For the application phase a molecule’s C-13 NMR spectrum and empirical formula must be available. Using the empirical formula, any compatible substructures present in the matrix are identified. These can be used to build a list of potential substructures giving rise to the experimental spectrum being analysed. Thus, the chemical space is drastically reduced. A Perl script has been written which identifies carbon atoms in substructures and uses these as nodes to bolt building blocks together into larger substructures.

In order to deduce which of the built isomers actually relates to the experimental spectrum, each must be put through the same substructural analysis phase previously described. By breaking down each isomer into its substructures, these can be compared to frequencies within the substructural matrix. The frequencies are treated as probabilities of substructures being present in a molecule, depending on chemical shift ranges. The ACO program also needs to respond to the fact that groups of substructures occurring together distort typical ranges, because of the secondary magnetic field effects.

The ACO can generate a shortlist of probable molecules, but to further reduce the search space, a spectrum needs to be predicted for each one. Each predicted spectra then needs to be aligned to the experimental spectrum and ranked accordingly. It can be difficult to determine corresponding shifts between predicted and experimental spectra, especially

where multiple shifts occur within a small separation. A recent study has highlighted how matrices can be used to detect optimal matches between experimental and predicted spectra (Griffiths and Bright 2002). To overcome some of these issues, ranking will be decided via a string algorithm encoding substructural features, rather than chemical shifts. The highest matching isomer can then be taken as the most persuasive underlying structure.

Bibliography:

Bonabaeu, Eric, et al. (1999), “Swarm Intelligence: From Natural to Artificial Systems”; ISBN 0195131592 –Oxford University Press

Gasteiger, Johann (ed.) (2003), “Handbook of Chemoinformatics: From Data to Knowledge”; ISBN 3-527-30680-3 - Wiley-VCH, Weinheim

Griffiths, L. and J. D. Bright (2002). "Towards the automatic analysis of H-1 NMR spectra: Part 3. Confirmation of postulated chemical structure." Magnetic Resonance in Chemistry 40(10): 623-634.

Griffiths, L. (2003). "Automatic analysis of NMR spectra." Annual Reports on NMR Spectroscopy 50: 217-251.

Hornak, J. P. (1997). The Basics of NMR.