Issues in aligning multiple -MS spectra

David Clifford*¹, Glenn Stone¹, David Lovell¹, Ziad Ramadan², & Sunil Kochhar²

¹CSIRO Mathematical and Information Sciences, Australia
²The Nestlé Research Centre, Lausanne, Switzerland

1 Introduction

Mass spectrometry (MS) is a popular means for measuring the abundance and mass/charge ratio of molecules in a sample. MS is often preceded by an elution process that separates different molecular species based on characteristics like molecular size or hydrophobicity. Gas chromatography (GC) or liquid chromatography (LC) are two common elution methods and the range of elution and mass spectrometry combinations are referred to as “hyphenated MS” (e.g. GC-MS, LC-MS etc).

The elution introduces variation in the time that different molecular species enter the mass spectrometer. While on any particular occasion the elution profiles will be consistent, over the course of several months there is a noticeable difference in elution profiles. Thus, one of the first steps in analysing “hyphenated MS” data is to align the spectra measured from the different samples under study.

We have a dataset of 712 GC-MS recordings of human plasma, recorded during a time course study and processed in several batches over the course of three months. The data come from the NUGENOB project, an EU initiated project (number QLK1-CT-2000-00618) designed to elucidate the role, in human obesity, of interactions between macronutrient compositions of the diet with particular emphasis on fat intake. In its most raw format, each of the GC-MS datafiles is a matrix of non-negative (integer) ion intensity values of (approximate) dimension 36000 × 750.

The MS typically used nowadays is precise and no alignment is needed in the mass-charge dimension once the machine is properly calibrated. The chromatogram for a GC-MS file is the total ion count (TIC) recorded by the mass-spectrometer at a particular time; often the alignment is based on the TIC, ignoring information recorded in the MS dimension.

Alignment is not a trivial exercise given the amount of data involved and the variation that the elution process can cause. The computational problem lends itself to parallelisation and high performance computing.

2 Need for alignment

Figure 1A shows part of the chromatograms for each file from a single batch of GC-MS data. Compare this to Figure 1B where we plot the average chromatograms from three different batches. There is a clear need for alignment of files from different batches before one can begin to look for differences between groups in the data. A linear shift or rescaling of the elution times from one batch the the next does not align all the peaks. A more sophisticated warping of the elution times is required to correct for the drift in the GC.
Figure 1: A. The top image shows raw chromatograms from a single batch. Within any batch there is little need for an alignment. B. The middle image shows the average chromatogram from three different batches (including the batch shown in the top image, plotted here in black). C. The lowest image shows the same chromatograms after penalised DTW.
3 Methods for alignment

Two methods appear to be well suited to this alignment problem: dynamic time warping (DTW) and robust point matching (RPM).

DTW (Wang & Isenhour, 1987) is a computationally efficient method of local feature alignment. DTW repeats or removes intensity values in a pair of signals to minimise the distance between them. This can lead to artificial features in aligned signals, such as wide flat peaks or abrupt changes in intensity. The number of artificial features can be reduced using penalties. DTW can be extended to work with the MS dimension of the data.

RPM (Saussen, 2007; Kirchner et al., 2007) typically uses smooth monotone regression (Ramsay & Silverman, 2005) to align features. RPM works on chromatograms but also on GC-MS or LC-MS images themselves once a set of peaks has been chosen. RPM is computationally slow even for chromatograms of length 1000 and its use with multiple chromatograms of length 36000 is not currently feasible. RPM is asymmetric, aligning signal A to signal B, and aligning B to A can give very different results. A version of RPM that uses monotone kernel regression (Pelckmans et al., 2005) is currently under development within the amsrpm package and appears to be more computationally efficient (Saussen, 2007).

4 Conclusions

RPM, in its current form is not feasible given the length of the chromatograms we wish to align. We focus on DTW and improve its performance with penalty terms and a metric that uses MS information. This increases the time needed for evaluating the DTW but allows us to incorporate more information and results in a better alignment overall. We examine the average chromatogram from different batches and align them, see Figure 1C.

References


