Improving Untargeted Differential Analysis of Mass Spectrometric Data by Recursive Feature Extraction

Joseph Roark; Norton Kitagawa; Steven M. Fischer; Theodore Sana; David Peterson; Ed Darland; Xiangdong Li - Agilent Technologies, Inc., Santa Clara, CA

IMSC 2009

Agilent Technologies

Introduction

Processing of un-targeted mass spectrometric data for chemometric analysis involves a feature extraction step, followed by an alignment of features across multiple sample data sets. Feature extraction algorithms for LC/MS and GC/MS data often process single data files at a time, owing to the complexity and memory requirements of the raw data. Due to the high cost in processing time for large data sets, most feature extraction algorithms are optimized for speed, at the expense of sensitivity and quality, which can introduce both false positives and false negatives. Additionally, alignment of data containing false positives results in a sparse matrix of mass-retention time-abundance values (or mass spectrum-retention time-abundance in GC/MS), which can be challenging to analyze statistically. These deficiencies in feature extraction result in unreliable post-processing statistics and introduce noise into abundance profiling experiments.

The "missing data problem" is not unique to the analysis of mass spectrometric data. Various imputation techniques for replacing missing data have been developed for use in psychology (Graham 2009), epidemiology (Donders et al. 2006) and microarray data (Sehgal 2009). We intend to demonstrate, however, that there is great benefit in recursively mining the original data set, which simultaneously removes false positives by validating each feature and extracts a non-zero abundance value for features that were missed in the first pass, due to low abundance, low signal/noise, etc. We have developed a two-pass recursive feature extraction workflow and algorithms to significantly improve the mass spectrometric feature extraction on large data sets.

Experimental

Initial Feature Extraction

24 LCMS data files were divided up into 6 experimental conditions with 4 technical replicates. Compound features were initially extracted using a proprietary algorithm called Molecular Feature Extractor (MFE) in the MassHunter Qualitative Analysis (MassHunter Qual) software. Compounds with a total abundance of less than 5000 counts were filtered away from the initial set. Extracted features for each LC/MS data file were written out to XML-based Compound Exchange Format (.CEF) files and imported into MassHunter Mass Profiler Professional (MPP) for chemometric analysis.

Binning and Alignment

Compound features were binned and aligned across 24 sample files using a nearest neighbor search algorithm based on neutral mass and retention time values using tight tolerances. The result is a composite feature list across all individual samples. The population density of each compound bin were calculated and used to partition the candidate compound lists for subsequent re-extraction.

Binning Window	Relative Tol.	Fixed Tol.
Mass	5 ppm	2 mDa
Retention Time	0.1%	0.15 min

Recursive Feature Extraction

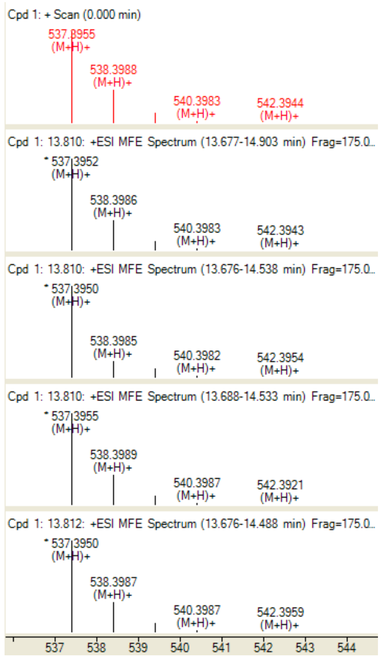
The binned composite compound feature list was used by a second algorithm called Find by Ion (new in MassHunter Qual). Each compound abundance was re-determined by extracting a chromatogram from all ions associated with that compound (EIC) in each sample in the original data set and the chromatogram was integrated. Chromatograms that failed to integrate an EIC peak at the originally measured retention time were rejected as being invalid. Re-extracted compounds were exported to a CEF file and imported into MPP for a second pass statistical analysis. The two feature extraction sets were subsequently co-analyzed in MPP for overlap and bin density. A few representative features were also checked for retention time, m/z and abundance deviation.

Composite Spectrum Creation

For each feature in the binned composite list, a composite spectrum of all found ions was created across all samples in the experiment. This composite spectrum is used as input into the recursive feature extraction algorithm. Doing so allows the algorithm to extract EIC's for any and all found ions, as well as provide a more representative observed spectrum and observed isotope ratio. The composite spectrum is also used for database searching and empirical formula calculation.

Results and Discussion

Figure 1. Composite spectrum. The bottom four plots in black are the same feature (536.388 Da at 13.808 min), detected in only one sample group in a differential analysis. The composite spectrum is shown in red.



Recursive Feature Extraction

Using the composite spectrum created in Figure 1, the recursive feature extraction algorithm extracted EIC's, using the m/z values provided by the composite spectrum and a +/- 1.5 minute retention time window. Providing this limited extraction window improves performance by reducing the search space, as well as prevents erroneous identification of isomers (which are detected as separate features).

Figure 2. Recursive extraction found the 536.388 Da at 13.808 min in both sample groups instead of only one group. Recursively extracted feature shown in green.

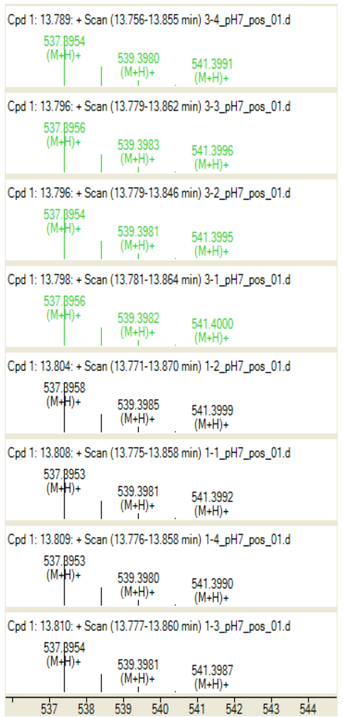
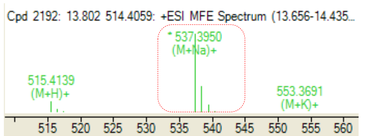


Figure 3. Closer inspection revealed that the feature was missed the first time due to the ion being included as an $[M+Na]^+$ isotope cluster with a neutral mass of 514.406 Da. Both putative features were extracted recursively.



Results and Discussion

Figure 4. An example of a feature found in one sample group, but missed completely within another sample group by MFE. After recursion, the feature was recovered in the empty group. (A) EIC's from four sample files. (B) Isotope pattern of $[M+H]^+$ ion.

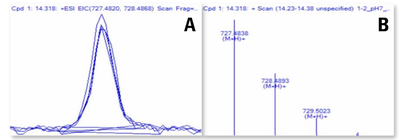
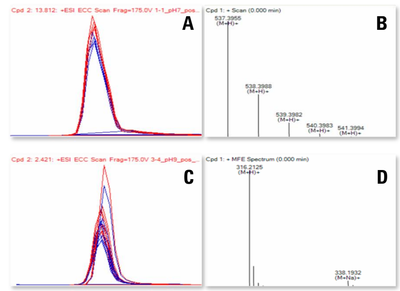


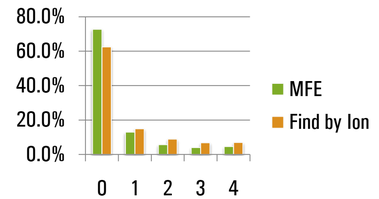
Figure 5. Two examples of features found by MFE and validated using recursion. (A) and (C) are EIC's for two independent features. (B) and (D) are their respective composite spectra for two features. MFE in red, recursive extracted in blue.



Bin Population

A decrease in the number of empty bins is indicative of the recursive feature extraction finding more features where none were detected before. The overall increase in bin population is indicative of bins containing missing feature that have now been detected by recursive feature extraction.

Figure 6. Histogram of bin population across the 24 LCMS data files. Graph displays percentage of feature bins versus the number of non-zero abundance values per bin.



Conclusions

- Feature extraction algorithms are typically designed to process one data file at a time.
- Exhaustive feature extraction is time and resource consuming and requires substantial data trade-offs.
- We propose a new two-pass recursive feature finding approach, in which
  - An initial feature extraction is performed per data file using the fast MFE algorithm.
  - Features are binned and aligned into a composite feature list across all samples and a composite spectrum is created for each feature.
  - The composite feature list is re-extracted using a different algorithm in a targeted mode.
- Recursive feature extraction improves the quality of mass spectrometric feature data by subjecting features to validation (both false positives and false negatives)
- Though not shown here, the recursive approach may be applied to GCMS data as well.
- Future work will focus on further improving the quality of features in the initial Molecular Feature Extraction and performance improvements on the recursive feature extraction.

References

1. Graham JW. Annu. Rev. Pyschol. 2009, 60:549-576
2. Donders ART, van der Heijden G, Stijnen T and Moons K. J. Clin. Epidemiol. 2006, 1087-1091.
3. Sehgal MSB, Gondal I, Dooley LS and Coppel R. EURASIP J. Bioinform. Syst. Biol. 2009, 717136