Genome analysis

SpecAlign—processing and alignment of mass spectra datasets

Jason W. H. Wong1,*, Gerard Cagney2 and Hugh M. Cartwright1

1Chemistry Department, Oxford University, Physical and Theoretical Chemistry Laboratory, South Parks Road, Oxford OX1 3QZ, UK and 2Conway Institute, University College Dublin, Belfield, Dublin 4, Ireland

Received on December 10, 2004; revised on January 19, 2005; accepted on January 27, 2005
Advance Access publication February 2, 2005

ABSTRACT

Summary: Pre-processing of chromatographic profile or mass spectral data is an important aspect of many types of proteomics and biomarker discovery experiments. Here we present a graphical computational tool, SpecAlign, that enables simultaneous visualization and manipulation of multiple datasets. SpecAlign not only provides all common processing functions, but also uniquely implements an algorithm that enables the complete alignment of each mass spectrum within a loaded dataset. We demonstrate its utility by aligning two datasets each containing six spectra; one set was acquired prior to instrument calibration and the other following calibration.

Availability: The software is free of charge and available for download from http://ptcl.chem.ox.ac.uk/~jwong specalign. Supports Windows operating systems including Windows 9X/NT/2000/XP.

Contact: jason.wong@chem.ox.ac.uk

INTRODUCTION

Proteomics aims in a single experiment to describe the identity and relative abundance of large numbers of proteins. In recent years, such efforts have led to the development of several ‘profiling’ technologies whereby chromatographic and mass spectral datasets may be generated from numerous biological samples. The rapid spread of these technologies has generated very large spectral datasets that need to be compared and analyzed. However the tools and skills needed to do this are often unavailable, or are available only for a specific instrument. A simple-to-use research tool suitable for handling different types of proteomics spectral data independent of the technology platform used to obtain it would be very beneficial. Several types of chromatographic and spectral datasets are currently encountered in proteomics laboratories. Files of peaks representing separated biomolecules are often generated as the preliminary step of a proteomics experiment (before or during mass spectrometry) and may be obtained using UV or ion detectors (e.g. total ion chromatograms, TIC). Using mass spectrometry, spectral files representing intact or fragmented proteins and peptides can be generated in experiments that often are of a very large scale (Aebersold and Mann, 2003; Yates, 2004). Examples are diverse and include spectral profiles of protein peaks obtained from multiple tissue samples using matrix-assisted laser desorption ionization (MALDI) or the related method of surface-enhanced laser desorption ionization (SELDI) mass spectrometry. These datasets represent expression profiles of the relevant tissues (the mass of the proteins and their relative expression being represented on the x and y axes) and are being used to search for biomarkers that may serve as early indicators of disease (Diamandis, 2004). A very common proteomics method is to digest proteins into tryptic fragments that generate patterns upon MALDI analysis (so-called ‘peptide mass fingerprints’) that can be used to search protein sequence databases. Similarly, spectral alignment and averaging is useful for the analysis of tandem mass spectra of peptides. In this case, peptide ions isolated in mass spectrometers are ‘fragmented’ into daughter ion patterns that can be used to infer amino acid sequence information by comparison with sequence patterns predicted using protein sequence databases (Nesvizhskii and Aebersold, 2004).

However, a common feature of these approaches is the need for easy-to-use applications that can handle multiple large datasets. For instance, a single SELDI–MS dataset may encapsulate very complex clinical proteomics sample data comprising tens of thousands of mass positions and associated intensity values; hence analysis is rarely straightforward and usually requires substantial pre-processing before the data can be further analyzed by statistical or machine learning methods. In particular, instrument resolution or instrument calibration may affect the quality of datasets (e.g. for SELDI the variance may be ±0.1–0.2% of the mass/charge ratio at any point; Yasui et al., 2003); therefore alignment of spectra within datasets is often required.

Peak alignment tools are available in commercial applications, but normally these are specific to particular instruments or applications, with input and output formats that are difficult to integrate with upstream or downstream analysis. Therefore, the development of SpecAlign is motivated by the need for a tool that enables the alignment of complete mass spectra. A number of algorithms exist for the alignment of spectral data (Torgrip et al., 2003), but as far as we know, none are implemented as a widely accessible software tool. We briefly describe the alignment algorithm and program operation in the following sections. Instructions for using the features of SpecAlign are discussed fully at http://ptcl.chem.ox.ac.uk/~jwong specalign/support.htm

ALIGNMENT ALGORITHM

The spectral alignment algorithm implemented is unique to SpecAlign. It is designed to enable the alignment of two or more mass
spectra, each of which may contain tens of thousands of data points, within a short period of time (<1 min) on a standard personal computer. A heuristic algorithm was developed that has a computational complexity of $O(ds)$, where $d$ is the number of spectral data points and $s$ is the number of spectra. This algorithm is based on the insertion and deletion of data points to shift regions in each spectrum, $m$, to align with the corresponding region in a reference spectrum, $r$, as marked by reference points, $Pim$ and $Pjr$, where $i$ and $j$ are points between 0 and $d$. By default, the algorithm makes use of an average spectrum (comprising all spectra to be aligned) as a reference, although a user-specified spectrum may be used. Reference points typically consist of automatically selected peaks, but may also consist of manually selected peaks or points within each spectrum.

The algorithm proceeds as follows for each spectrum, $m$, to be aligned to the reference spectrum, $r$:

1. For each $j$ in $Pjr$, find the closest matching $Pim$. If no match is found within a window of a size, $w$, specified by the user, then move to the next point $j + 1$.

2. If $Pim$ is found but not aligned to $Pjr$, find the minima between $Pim$ and $P(i-1)m$, $P(i-1)m$, $min_{-1}$ and $P(i+1)m$, $P(i+1)m$, $min_{+1}$, where insertions or deletions are to be made for alignment of $Pim$ to $Pjr$.

3. If $Pim > Pjr$ (for the value of the x-axis), then points are to be deleted from the $min_{-1}$ and points to be inserted at $min_{+1}$. If $Pim < Pjr$ then the reverse applies.
(4) Where points are inserted, the y-axis value for the inserted point is estimated by a least squares quadratic polynomial fit to its adjacent w points.

In theory, information may be lost at points of insertion and deletion; however for applications such as mass spectral data analysis for biomarker discovery, there should be little impact as signals in mass spectrometry are only ever represented by peaks and not as minima or troughs.

Figure 1 shows an example of an alignment of a MALDI mass spectral dataset containing both samples acquired before and after instrument calibration. It can be seen that before alignment, the dotted lines representing spectra acquired from an uncalibrated instrument are poorly aligned to those acquired after the instrument was calibrated. Following alignment by the algorithm described above, peaks from each spectrum become aligned, enabling more accurate comparisons to be made between all spectra. A general example of the advantage of mass spectral alignment is demonstrated in tandem mass spectrometry database searching by Pevzner et al. (2001).

PROGRAM DESCRIPTION
SpecAlign has been implemented using C++, using the Microsoft Foundation Classes libraries for the development of the graphic user interface. Users may import spectral data files of any type as ASCII comma delimited or tab delimited files, where the first column represents the x-axis and the second represents the y-axis. Once data files have been loaded, users may interactively zoom in/out, crop, select/remove peaks for all spectra simultaneously. The spectra may be viewed as a line graph, bar graph or all stacked on one axis. SpecAlign also provides spectral processing tools including normalization by total spectrum signal, conversion to relative intensities, subtraction of baseline, scaling about the y-axis to enhance small peaks or to suppress noise, smoothing by the Savitzky–Golay filter (Savitzky and Golay, 1964), binning values about the x-axis, automatically picking peaks based on default or user-defined parameters, and finally spectral alignment as described in the previous section. All processing methods are designed with the principle aim of rendering spectral datasets ready for further analysis by statistical or machine learning methods. Consequently, SpecAlign provides methods to export any processed data to ASCII comma delimited files. Finally, users may also save any processed data in SpecAlign’s native format (file extension, SPA) for convenience of data storage and exchange.

Any type of chromatographic or spectral data may be visualized and processed as described.

CONCLUSION
With SpecAlign, a tool has been created for the visualization and manipulation of multiple mass spectral datasets to address challenges in proteomic data analysis. Most significantly it enables researchers to rapidly align spectral datasets for further analysis by other methods. As the underlying algorithms to the processing method are readily comprehensible, researchers can have confidence in using SpecAlign in the analysis and processing of their data.

REFERENCES