washAlign: a GC-MS Data Alignment Tool Using Iterative Block-Shifting of Peak Retention Times Based on Mass-Spectral Data

Minho Chae[1, 3], John J. Thaden[3, 5], Steven F. Jennings[2], and Robert J. Shmookler Reis[3, 4, 5]

[1]UALR/UAMS Joint Graduate Program in Bioinformatics, University of Arkansas at Little Rock, Little Rock, AR 72204
[2]Department of Information Science, University of Arkansas at Little Rock, Little Rock, AR 72204
[3]Department of Geriatrics, and [4]Department of Biochemistry and Molecular Biology, University of Arkansas for Medical Sciences, Little Rock, AR 72205
[5]Central Arkansas Veterans Healthcare System LRVA-151, 4300 W. 7th Street, Little Rock, AR 72205

Email addresses: MC: mxchae@ualr.edu, JT: jthaden@uams.edu, SJ: sfjennings@ualr.edu, and RJSR: rjsr@uams.edu

In GC-MS, a gas chromatograph (GC) resolves chemicals by time of elution from a coated capillary through which gas flows; a mass spectrometer (MS) resolves ions (produced upon fragmentation of eluates) by mass/charge (m/z) ratio; and an acquisition program records ion intensity as a function of m/z and elution, yielding spectra and chromatograms, respectively. A problem when comparing records in an experiment is that elution times will vary. washAlign has been developed in R to address this problem. It warps regions between peaks that it has shifted, thereby aligning those peaks to spectrally matched peaks in a reference chromatogram while preserving their shape and area. Through pair-wise comparisons of all records to one arbitrarily selected reference record, all records in a large experiment can be aligned for subsequent processing, *e.g.*, by three-way methods, including those such as PARAFAC that assume mathematical trilinearity.

In washAlign, (a) ion chromatograms are extracted for a subset of those m/z channels with the five highest ion intensities in any of the consecutive MS scans that define "a region of the sample and reference chromatograms that exhibit a peak on the total intensity chromatogram"; (b) peaks are detected in them, and key peaks are matched between sample and reference through a procedure involving iterative localization with spectral correlation, to produce for each sample and the reference a peak list for alignment; and (c) the key sample peaks are shifted toward the matching peaks in the reference run, and nonpeak regions are warped, i.e., linearly interpolated, to join the shifted peak regions.

Users can visually inspect the chromatograms before and after alignment of a pair of chromatograms, through an interactive selection of matched peaks. Taking an iterative block-shift approach makes it possible to not only reveal strongly matching peaks at early stages but also to reduce the risk of mismatching chemically different peaks.