Using the MassSpecWavelet package

Pan Du[‡,*], Warren A. Kibbe[‡,†], Simon Lin[‡,‡]

May 3, 2016

[‡]Robert H. Lurie Comprehensive Cancer Center
Northwestern University, Chicago, IL, 60611, USA

Contents

1 Overview of MassSpecWavelet

MassSpecWavelet R package is aimed to process Mass Spectrometry (MS) data mainly based on Wavelet Transforms. The current version only supports the peak detection based on Continuous Wavelet Transform (CWT). More functions covering baseline removal, smoothing, alignment will be added in the future versions. The algorithms have been evaluated with low resolution mass spectra (SELDI and MALDI data), we believe some of the algorithms can also be applied to other kind of spectra.

2 Citation

Du, P., Kibbe, W.A. and Lin, S.M. (2006) Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching, Bioinformatics, 22, 2059-2065.

3 Peak detection by using CWT-based pattern matching

Motivation: A major problem for current peak detection algorithms is that noise in Mass Spectrometry (MS) spectrum gives rise to a high rate of false positives. The false positive rate is especially problematic in detecting peaks with low amplitudes. Usually, various baseline correction algorithms and smoothing methods are applied before attempting peak detection. This approach is very sensitive to the amount of smoothing and aggressiveness of the baseline correction, which contribute to making peak detection results inconsistent between runs, instrumentation and analysis methods.

[*]dupan@northwestern.edu
[†]wakibbe@northwestern.edu
[‡]s-lin2@northwestern.edu

Results: Most peak detection algorithms simply identify peaks based on amplitude, ignoring the additional information present in the shape of the peaks in a spectrum. In our experience, 'true' peaks have characteristic shapes, and providing a shape-matching function that provides a 'goodness of fit' coefficient should provide a more robust peak identification method. Based on these observations, a Continuous Wavelet Transform (CWT)-based peak detection algorithm has been devised that identifies peaks with different scales and amplitudes. By transforming the spectrum into wavelet space, the pattern-matching problem is simplified and additionally provides a powerful technique for identifying and separating signal from spike noise and colored noise. This transformation, with the additional information provided by the 2-D CWT coefficients can greatly enhance the effective Signal-to-Noise Ratio (SNR). Furthermore, with this technique no baseline removal or peak smoothing preprocessing steps are required before peak detection, and this improves the robustness of peak detection under a variety of conditions. The algorithm was evaluated with real MS spectra with known polypeptide positions. Comparisons with two other popular algorithms were performed. The results show the CWT-based algorithm can identify both strong and weak peaks while keeping false positive rate low.

3.1 Continuous wavelet transform with Mexican Hat wavelet

Load the MassSpecWavelet library.

Load the example data

```
> data(exampleMS)
```

Continuous wavelet transform with Mexican Hat wavelet.
The 2-D CWT coefficients image of MS spectrum in [5000, 11000] is shown in Figure **??**

```
> scales <- seq(1, 64, 2)
>  wCoefs <- cwt(exampleMS, scales = scales, wavelet = "mexh")
```

3.2 Peak identification process

Identify the ridges by linking the local maxima
The identified local maxima is shown in Figure **??**

```
> ## Attach the raw spectrum as the first column
> wCoefs <- cbind(as.vector(exampleMS), wCoefs)
> colnames(wCoefs) <- c(0, scales)
> localMax <- getLocalMaximumCWT(wCoefs)
>
```

Identify the ridge lines by connecting local maxima of CWT coefficient at adjacent scales

```
> ridgeList <- getRidge(localMax)
```

Identify the identified ridges lines and SNR using `identifyMajorPeaks`. The returns of `identifyMajorPeaks` include the peakIndex, peakSNR and etc. All these elements carry peak names, which are the same as the corresponding peak ridges. See function `getRidge` for details.

```
> ## Plot the 2-D CWT coefficients as image (It may take a while!)
> xTickInterval <- 1000
> plotRange <- c(5000, 11000)
> image(plotRange[1]:plotRange[2], scales, wCoefs[plotRange[1]:plotRange[2],], col=terrain
> axis(1, at=seq(plotRange[1], plotRange[2], by=xTickInterval))
> axis(2, at=c(1, seq(10, 64, by=10)))
> box()
```

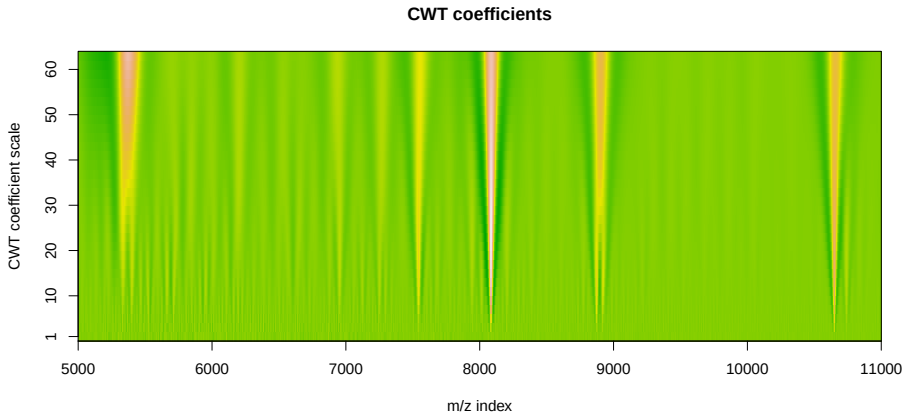

Figure 1: 2-D CWT coefficient image

```
> plotLocalMax(localMax, wCoefs, range=plotRange)
```

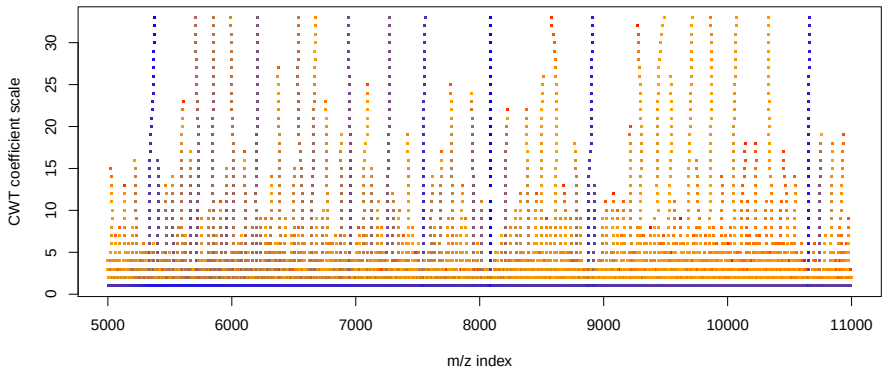

Figure 2: Identified local maxima of CWT coefficients at each scale

```
> plotRidgeList(ridgeList,  wCoefs, range=plotRange)
```

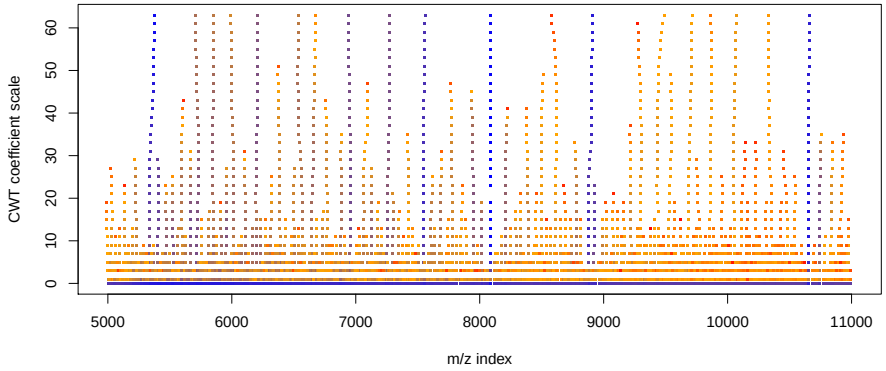

Figure 3: Identified ridge lines based on 2-D CWT coefficients

```
> SNR.Th <- 3
> nearbyPeak <- TRUE
> majorPeakInfo <- identifyMajorPeaks(exampleMS, ridgeList, wCoefs, SNR.Th = SNR.Th, nearb
> ## Plot the identified peaks
> peakIndex <- majorPeakInfo$peakIndex
```

Plot the specta with identified peaks marked with read circles.

All of the above steps are encapsulated as a main function of peak detection main

```
> data(exampleMS)
> SNR.Th <- 3
> nearbyPeak <- TRUE
> peakInfo <- peakDetectionCWT(exampleMS, SNR.Th=SNR.Th, nearbyPeak=nearbyPeak)
> majorPeakInfo = peakInfo$majorPeakInfo
> peakIndex <- majorPeakInfo$peakIndex
> plotRange <- c(5000, length(exampleMS))
```

Plot Signal to Noise Ration (SNR) of the peaks

```
>         peakSNR <- majorPeakInfo$peakSNR
>         allPeakIndex <- majorPeakInfo$allPeakIndex
```

3.3 Refine the peak parameter estimation

The above peak detection process can identify the peaks, however, it can only approximately estimate the peak parameters, like peak strength (proportional to Area Under Curve), peak center position and peak width. In order to get better estimation of these parameter, an estimation refine step can be added.

```
>         betterPeakInfo <- tuneInPeakInfo(exampleMS, majorPeakInfo)
```

```
> plotPeak(exampleMS, peakIndex, range=plotRange, main=paste('Identified peaks with SNR >'
```

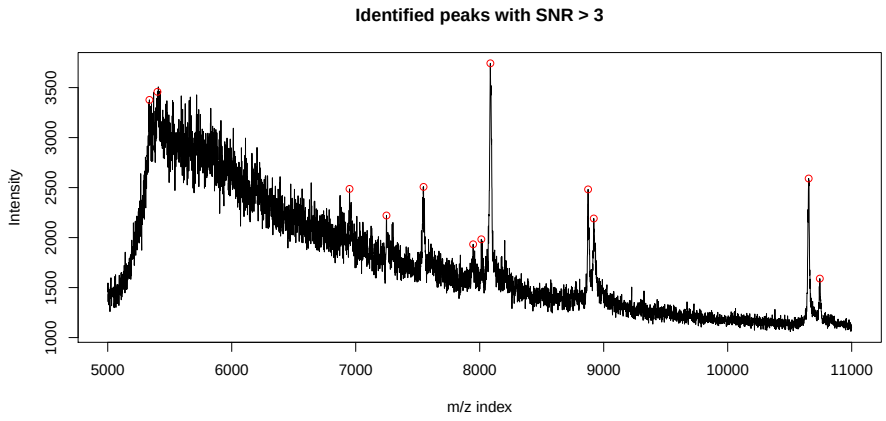

Figure 4: Identified peaks

```
> plotPeak(exampleMS, peakIndex, range=plotRange, log='x', main=paste('Identified peaks wi
```

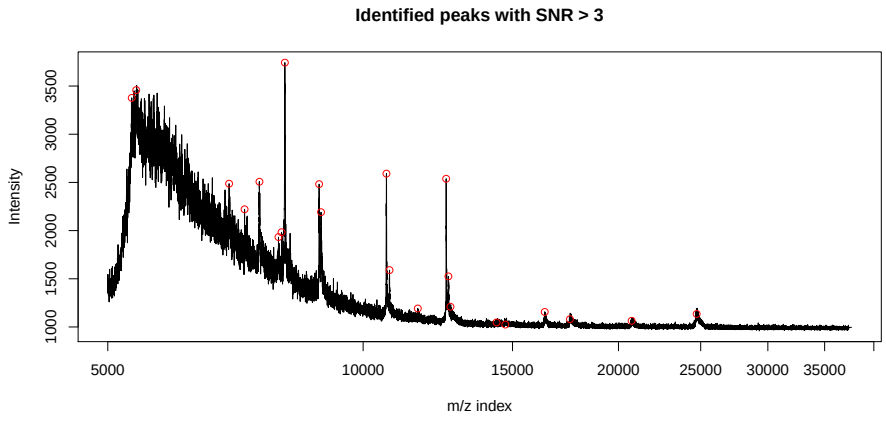

Figure 5: Identified peaks

```
>         plotRange <- c(5000, 36000)
>         selInd <- which(allPeakIndex >= plotRange[1] & allPeakIndex < plotRange[2])
>         plot(allPeakIndex[selInd], peakSNR[selInd], type='h', xlab='m/z Index', ylab='Si
>         points(peakIndex, peakSNR[names(peakIndex)], type='h', col='red')
>         title('Signal to Noise Ratio (SNR) of the peaks (CWT method)')
```

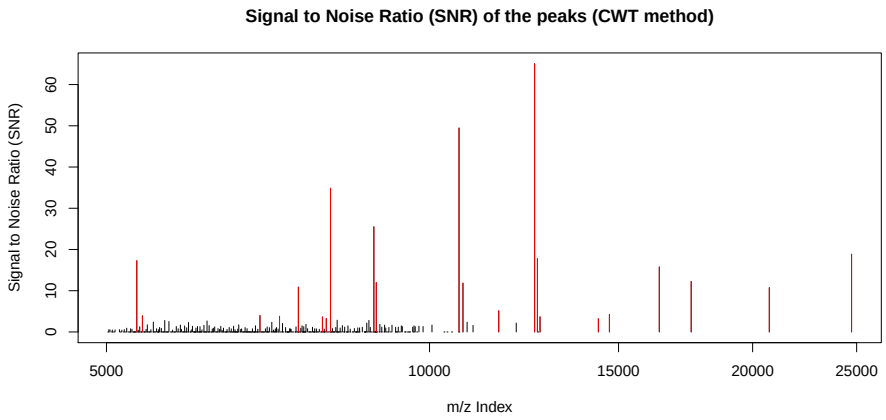

Figure 6: Estimated Signal to Noise Ration (SNR) of the peaks

```
>         plotRange <- c(5000, 11000)
>                 plot(plotRange[1]:plotRange[2], exampleMS[plotRange[1]:plotRange[2]], ty
>                 abline(v=betterPeakInfo$peakCenterIndex, col='red')
```

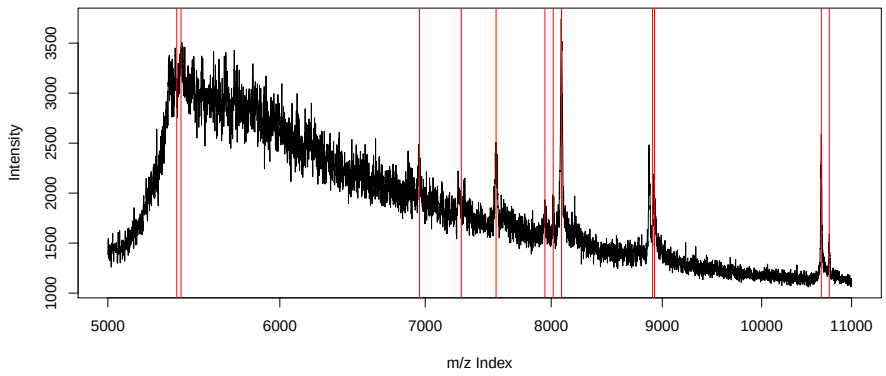

Figure 7: Identified peaks with refined peak center position

4 Future Extension

More MS data analysis and wavelet related functions will be implemented in MassSpecWavelet package.

5 Acknowledgments

We would like to thanks the users and researchers around the world contribute to the MassSpecWavelet package, provide great comments and suggestions and report bugs. Especially, we would like to thanks Steffen Neumann and Ralf Tautenhahn fixing some bugs for the package.