Application of dual tree complex wavelet transform in tandem mass spectrometry

Selvaraju Murugesan a, David B.H. Tay a,*, Ira Cooke b, Pierre Faou b

*Department of Electronic Engineering, La Trobe University, Bundoora, Victoria 3086, Australia
bDepartment of Biochemistry, La Trobe University, Bundoora, Victoria 3086, Australia

ARTICLE INFO

Article history:
Received 8 February 2015
Accepted 2 May 2015

Keywords:
Tandem mass spectrometry
Proteomic data processing
Dual tree complex wavelet transform
Signal denoising
Peptides detection

ABSTRACT

Mass Spectrometry (MS) is a widely used technique in molecular biology for high throughput identification and sequencing of peptides (and proteins). Tandem mass spectrometry (MS/MS) is a specialised mass spectrometry technique whereby the sequence of peptides can be determined. Preprocessing of the MS/MS data is indispensable before performing any statistical analysis on the data. In this work, preprocessing of MS/MS data is proposed based on the Dual Tree Complex Wavelet Transform (DTCWT) using almost symmetric Hilbert pair of wavelets. After the preprocessing step, the identification of peptides is done using the database search approach. The performance of the proposed preprocessing technique is evaluated by comparing its performance against Discrete Wavelet Transform (DWT) and Stationary Wavelet Transform (SWT). The preprocessing performed using DTCWT identified more peptides compared to DWT and SWT.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

Many techniques in analytical chemistry involve matching observed signals against an expected theoretical pattern. Examples include identification of analytes based on chromatographic signatures, analysis of 2D gel electrophoretic images and identification of peptides by tandem mass spectrometry. Selection of appropriate signal processing algorithms can significantly affect the sensitivity and reproducibility of end results when using such techniques [1,2]. This is particularly important in the field of mass spectrometry based proteomics, where recent advancements have led to a rapid expansion of the volume of data that is available for processing, and where novel approaches such as SWATH [3] place greater demands on the data.

Mass Spectrometry (MS) based proteomic analysis is a widely used technique to study interrelations between protein expressions and also to study relationships between proteins themselves [4]. The basic principle behind MS in proteomics is to fragment complex protein molecules via soft-ionisation techniques into smaller molecules such as peptides or amino acids so they are more readily analyzed [6]. The fragmented ions are separated according to their mass to charge ratio (m/z) which is measured in Daltons. A typical MS scan of a known dataset 3666 is shown in Fig. 1. Peptide mass fingerprint is determined by the extraction of the set of measured peptide masses. There are many algorithms developed [7,8] that match the experimental data against the theoretical masses obtained from the in silico digestion at the same enzyme cleavage sites of all protein amino acid sequences in the database. The proteins in the database are then ranked according to the number of peptide masses matching their sequence within a given mass error tolerance. Tandem mass spectrometry is a specialised mass spectrometry technique whereby the sequence of peptides can be determined. The precise identification of peptides and proteins is indispensable in developing new drugs for the treatment of human diseases such as cancer, diabetes and asthma. Tandem mass spectrometry is also called the MS/MS technique. In the MS/MS technique, peptide ions of interest are first selected in a precursor ion scan. Those ions selection is based on relative abundance. The direct derivation of peptide sequence from the MS/MS spectrum can be obtained by matching to theoretical spectra from peptides in a database (sequence database search) [9], matching to curated reference spectra (spectral database search) [10], or de novo sequencing [11]. The method in most widespread use is sequence database search because existing spectral databases are not comprehensive enough, and because de novo sequencing is often impractical due to the existence of missing peaks in most MS/MS spectra. In the sequence database search method the m/z of the precursor ion is used to select a set of candidate peptides with matching masses.
from a database. Theoretical MS/MS spectra are then generated for these candidate peptides using the rules of peptide fragmentation, and these are compared with the experimental MS/MS spectrum, and the best match is determined using a predefined scoring system. The MS/MS spectrum is usually corrupted by electrical noise, chemical noise and machine artifacts. Performing statistical and computational analysis on the noisy MS/MS is an extremely challenging task. Thus preprocessing of MS/MS data plays a vital role in identification of peptides. If the preprocessing of MS/MS produces false peaks, it will reduce the number of correctly identified peptides in the sample.

The Discrete Wavelet Transform (DWT) [12,13] is a versatile tool that has been used in a plethora of applications [14–17]. Wavelets are also widely used in the preprocessing stages of the proteomics data as it facilitates multi-resolution analysis [18,19]. The isotope wavelet transform proposed in [20] is an efficient framework for detecting isoipe patterns in the MS data sets. The isotope wavelet transform shows the versatility of the wavelet transform in detecting region of interest in the MS scans with low number of false positives. Coombes et al. [21] used the undecimated discrete wavelet transform (DWT) to denoise the MS spectra. However the undecimated DWT requires higher computational time compared to the critically decimated DWT. The rationale behind using the undecimated discrete wavelet transform is that it is shift invariant. Kwon et al. [22] used wavelet based denoising technique to reduce the noise from the MS data. Kwon et al. [22] used the undecimated DWT to remove the chemical and instrument noise from MS spectra using the hard thresholding technique. Kwon et al. [22] showed that the noise is heterogeneous in the MS experiments and it is not uniform in each MS scan. In their work, the data is segmented based on the variance change and the threshold for each of the segmented data is computed. The noise variance of each segment is estimated using the Median Absolute Deviation criterion [23]. To detect the variance change in the MS spectra, Kwon et al. [22] used an iterated cumulative sums of squares algorithm proposed by Gabbanini et al. [24]. Li et al. [25] showed that wavelet based denoising improved the performance of machine learning methods. Morris et al. [26] applied wavelet transform for feature extraction and quantification of MS data. The Dual Tree Complex Wavelet Transform (DTCWT) introduced by Kingsbury has emerged as one of the most popular redundant transform in a wide variety of applications [27–29]. The DTCWT has near shift invariance, provides directional selectivity in multidimensions and lower redundancy than the undecimated DWT [28]. In this work we present new techniques to preprocess the MS/MS using the Dual Tree Complex Wavelet Transform with the newly designed almost symmetric Hilbert pair of wavelets [30]. The previous works focus on the MS preprocessing and this is the first work to concentrate on the MS/MS preprocessing. The other novelty of the paper is the application of the DTCWT in preprocessing the tandem mass spectrometry data.

The overview of the paper is as follows. In Sections 2 and 3 we give a brief overview of the proteomics preprocessing stages and data collection process. The review of almost symmetric Hilbert pair of wavelets is presented in Section 4. In Section 5 we preprocess the MS/MS data using Dual Tree Complex Wavelet Transform and we discuss the results. The Section 6 concludes the paper.

2. Data collection

All the biological experiments for this paper were conducted at LaTrobe Institute of Molecular Science (LIMS), La Trobe University, Australia. We obtained six datasets that correspond to three biologically independent samples with two technical replicates per sample. All three biological samples correspond to extracts from human cancer cells after enriching for membrane glycoproteins and were subject to identical cell culture and sample preparation procedures. Importantly, all samples are relatively complex, containing many different peptides with different abundances. Prior to mass spectrometry all samples were reduced, alkylated and digested with trypsin. This process results in the production of a large number of distinct peptides of an appropriate size and charge for tandem mass spectrometry analysis. Peptide samples were then loaded onto a trap column (C18 PepMap 100 μm I.D. × 2 cm trapping column, Dionex) at 5 μL/min for 6 min and washed for 6 min before switching the precolumn in line with the analytical column (Vydac MS C18, 3 μm, 300 and 75 μm I.D. × 25 cm., Grace Pty. Ltd). The separation of peptides was performed at 300 mL/min using a linear ACN gradient of buffer A and buffer B (0.1% formic acid, 80% ACN), starting from 5% buffer B to 60% over 90 min. Data were collected on an hybrid quadrupole/time-of-flight MS (MicroTOF-Q, Bruker, Germany) with a nano-electrospray ion source using Data Dependent Acquisition mode and m/z 150–2500 as MS scan range. Nitrogen was used as the collision gas. The ionisation tip voltage and interface temperature were set at 4200 V and 205 °C, respectively. CID MS/MS spectra were collected for the 4 most intense ions. Dynamic exclusion parameters were set as follows: repeat count 2, duration 60 s. The data were collected and analysed using Data Analysis Software (Bruker Daltonics, Bremen, Germany).

3. Preprocessing of MS data

In order to identify and quantify the proteins in the sample, regions of interest corresponding to peptides (features) must be extracted from the raw MS spectra. These features consist of sets of closely spaced peaks whose arrangement can be used to deduce the peptide charge and monoisotopic mass. This information is combined with the MS/MS scans to determine the amino acid sequence of peptides. In addition, a quantitative measure for each feature is usually obtained by taking the area under the curve (AUC) across all the peaks associated with a feature, and this in turn can be used to infer the relative protein concentration.

Since the raw MS and MS/MS spectra are usually corrupted with noise and baseline artifacts, preprocessing plays a vital role in both quantitation and identification aspects of the experiment. In this paper we are mostly concerned with the application of the DTCWT algorithm to MS/MS spectra and its effects on the peptide identification process. In order to measure the effects of this algorithm on a single aspect of the system (MS/MS preprocessing) in isolation we used the msconvert tool [5] to apply standard
The first preprocessing step is noise filtering. The data from the MS experiment are corrupted by chemical and instrument noise. Since the volume of MS data is large, noise filtering is essential to clean the data. Wavelet based noise filtering is widely applied to denoise the MS data [22,21,20]. The hard thresholding of wavelet coefficients is performed after the decomposing the MS data using the DWT. Then the signal is reconstructed by taking inverse DWT. The baseline is estimated by taking the local minimum of wavelet functions, \( \phi(t) \) and \( \psi(t) \) respectively, are obtained from the filter coefficients via the two-scale equations \( \phi(t) = \sqrt{2}\sum_n h_0(n)\phi(2t-n) \) and \( \psi(t) = \sqrt{2}\sum_n h_1(n)\phi(2t-n) \). To ensure smooth scaling and wavelet functions zeros at \( z = -1 \) are imposed on \( H_0(z) \), and this is known as the (wavelet) vanishing moment (VM) condition. The DTCWT is implemented using a pair of two-channel perfect reconstruction (PR) multirate filter banks. The equivalent wavelet function of the two filter banks, \( \psi(t) \) and \( \psi(\tilde{t}) \) (with Fourier transform \( \Psi(\omega) \) and \( \Psi(\tilde{\omega}) \) respectively) should ideally satisfy the following Hilbert transform relationship:

\[
\Psi(\omega) = \begin{cases} -j\Psi(\omega), & \omega > 0 \\ \Psi(\omega), & \omega < 0 \end{cases}
\]

The wavelets \( \{\psi(t), \psi(\tilde{t})\} \) form a Hilbert-pair (and the same can be said of the corresponding pair of filter banks) and can either be biorthogonal or orthogonal. The dual-tree complex wavelet transform is based on a pair of filter banks. The upper and lower tree filters are denoted by superscripts \( h \) and \( g \) respectively. It is proven in [37] that Eq. (1) will hold if the low-pass filters of the filter banks, \( H_0(z) \) and \( H_2(z) \) satisfy

\[
H_0^*(e^{j\omega}) = e^{-j\omega/2}H_2^0(e^{j\omega}) \quad \text{for} \quad -\pi \leq \omega \leq \pi
\]

Thus the two filters \( H_0(z) \) and \( H_2(z) \) should satisfy the half sample delay condition (2). By defining \( \Psi(\omega) \) as the complex wavelet spectrum given by \( \Psi(\omega) = \Psi^R(\omega) + j\Psi^I(\omega) \), quality measures based on the level of analyticity can be defined [38] as

\[
E_1 = \frac{\max_{-\infty < \omega < 0} |\Psi^R(\omega)|}{\max_{-\infty < \omega < 0} |\Psi(\omega)|}, \quad E_2 = \int_{-\infty}^{0} \frac{\Psi^I(\omega)^2 d\omega}{|\Psi^R(\omega)|^2}
\]

Ideally if (1) is exact, then for \( \omega < 0 \), \( \Psi^I(\omega) = \Psi^R(\omega) + f(j\Psi^R(\omega)) = \Psi^R(\omega) - \Psi^I(\omega) = 0 \), i.e. \( \Psi(\omega) \) is complex analytic. This would mean \( E_1 = E_2 = 0 \) if (1) is exact. In practice (1) can only be approximated and therefore \( E_1 \) and \( E_2 \) are non-zero. The smaller the \( E_1 \) and \( E_2 \) values, the better the analytic quality. The complex valued (2) can be separated into magnitude and phase parts as follows:

\[
|H_0^R(\omega)| = |H_0^I(\omega)| \quad \text{(3)}
\]

\[
\Delta H_0^I(\omega) = -\omega/2 + \Delta H_0^R(\omega) \quad \text{(4)}
\]

The filters in the \( h \) and the \( g \) tree that are designed in the lattice framework by minimizing the measures \( E_{\text{sym}} \) and \( E_{\text{odd}} \)

\[
E_{\text{sym}} = \frac{1}{\frac{1}{n} \sum_{n=1}^L (h_0^R(n) - h_0^R(1-n))^2}
\]

\[
E_{\text{odd}} = \frac{1}{\frac{1}{n} \sum_{n=1}^L (h_0^R(n) - h_0^R(-n)) + (h_0^R(L))^2}
\]

The above conditions on the filters will satisfy the phase condition (4) exactly while the magnitude condition (3) is satisfied approximately. In [30], we have successively re-optimised the filters in \( g \) tree by relaxing the phase condition (4) such that a better magnitude approximation can be achieved to yield a better analytic quality Hilbert pair filters. The designed filters are almost symmetric orthogonal Hilbert pair of filters and they have approximately linear phase characteristics. We have designed length 12 Hilbert pair of filters having 2 vanishing moments [30]. Fig. 3 (top) shows the plot of \( \Psi(t) \), \( \Psi(\tilde{t}) \) and the corresponding complex envelope that are virtually symmetric. The complex wavelet spectrum shown in Fig. 3
using the DTCWT is described below:

$$y_f(t) = \hat{S}(t) + \hat{B}(t) + \epsilon(t)$$

where $y_f(t)$ is the observed signal, $s(t)$ is the original signal, $B(t)$ is the baseline drift and $\epsilon(t)$ is the noise which is modelled as zero mean white Gaussian. The main aim of the preprocessing stage is to get $s(t)$ from $y_f(t)$. The procedure to preprocess the MS/MS scans using the DTCWT is described below:

1. The first step is noise filtering. We observed from all our experiments that the noise in the MS/MS spectra does not show any variance change. We apply the DTCWT to decompose the MS/MS spectra. The hard thresholding technique is deployed to reduce the noise. The threshold ($T_p$) is set to three times the noise variance. The hard thresholding of the wavelet coefficient is defined as

$$\delta_j = \begin{cases} 
  d_{jk} & \text{if } |d_{jk}| > T_p \\
  0 & \text{if } |d_{jk}| \leq T_p 
\end{cases}$$

where $d_{jk}$ is the wavelet coefficient at the level $j$ at $k$th index. The noise variance is estimated using the Median Absolute Deviation [23] and it is given by

$$\sigma = \frac{\text{median}(d_{11})}{0.6745}$$

where $d_{11}$ is the detail coefficient at the first level.

2. After hard thresholding, we apply the inverse DTCWT to reconstruct the MS/MS signal. The next step is baseline correction. The baseline component is then removed by computing a monotone local minimum curve on the denoised signal. We have used the Matlab function `msbackadj` to correct the baseline and the window width for estimating the local minima is set to 200 separation units.

3. The signal is normalised by standardizing the area under the curve (AUC). The normalisation can be performed using the Matlab command `msnorm`.

4. Any local maximum after denoising, baseline correction and normalisation is assumed to be a peak. Firstly all the local maxima and the associated peak endpoints are computed. Then the signal to noise ratio at each local maxima is calculated. All the local maxima that are greater than a threshold is considered as peaks. The threshold is set at 10% of the maximum signal to noise ratio of that particular scan.

5. After peak picking, the peak lists are written into Mascot Generic Format (mgf) format and sent to the Mascot search engine [40]. The schema of the mgf format is available from Matrix Science, Ltd., London, United Kingdom. All Mascot searches used the following parameters: precursor mass tolerance 80 ppm, fragment ion mass tolerance 0.65Da, deamidated N as a variable modification, carbamidomethyl C as a fixed modification, up to one missed enzyme cleavage. The search database used included all Mammalian entries from RefSeq [36].

6. Raw Mascot search scores are then statistically analysed using Peptide Prophet [40] to produce a confidence level for each.

7. We then count the number of peptides obtained which have more than 95% of confidence level.

5.1. Discussion

The plot of first scan with a peptide match from dataset 3666 is shown in Fig. 4. This shows that most of the peaks are in the lower $m/z$ range and that this is also where most of the noise is located. Efficient preprocessing of the datasets will improve quality of MS/MS spectra and it will lead to a greater number of identified peptides. After pre- and post-processing (see Section 5) MS/MS spectra were submitted for database search using the Mascot search engine. This process attempts to match observed spectra with theoretical spectra from a large database of possible peptide sequences. We used the number of high scoring peptide sequence matches (PSM’s) as a measure of the performance of the preprocessing algorithm. The number of identified peptides is shown in Table 1.

We compare the results of DTCWT with other transforms such as DWT and Stationary Wavelet Transform (SWT) on the basis of the numbers of identified peptides. The length 12 Hilbert pair of filters having 2 VMs are used in DTCWT. The Daubechies length 12 filter having 6 VMs is used in DWT and SWT. In all the cases, the DTCWT performs better than the DWT and has a comparable performance to the SWT. However the computational load of the SWT is higher than the DTCWT.

6. Conclusion

In this paper, we have proposed a new preprocessing of the MS/MS data. We have evaluated the performance of three transforms namely DWT, SWT and DTCWT. We have shown that the DTCWT outperforms the DWT and SWT whereby detecting more number of peptides. The shift invariant properties of the DTCWT along with the approximate linear characteristics of the Hilbert pair.
of wavelets have proved to be very efficient compared to DWT and SWT.

Although our main focus has been on MS/MS data preprocessing it is also possible that the new preprocessing technique proposed here (DTCWT) will provide improvements to MS data processing. This would presumably lead to improvements in feature detection, charge assignment and estimates of peptide abundance. We believe that this would be a useful avenue for future work, but note that devising benchmarks is more challenging since key information derived from the MS signal such as feature intensity is very difficult to independently verify.

Conflict of interest

None declared.

References


Table 1

<table>
<thead>
<tr>
<th>Sample</th>
<th>Dataset</th>
<th>DWT</th>
<th>DTCWT</th>
<th>SWT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3666</td>
<td>59</td>
<td>64</td>
<td>61</td>
</tr>
<tr>
<td>1</td>
<td>3681</td>
<td>62</td>
<td>61</td>
<td>64</td>
</tr>
<tr>
<td>2</td>
<td>3670</td>
<td>59</td>
<td>67</td>
<td>67</td>
</tr>
<tr>
<td>2</td>
<td>3685</td>
<td>47</td>
<td>48</td>
<td>45</td>
</tr>
<tr>
<td>3</td>
<td>3677</td>
<td>63</td>
<td>63</td>
<td>62</td>
</tr>
<tr>
<td>3</td>
<td>3690</td>
<td>65</td>
<td>68</td>
<td>65</td>
</tr>
</tbody>
</table>


