Computers in Biology and Medicine 63 (2015) 36–41

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

Application of dual tree complex wavelet transform in tandem mass spectrometry

Selvaraaju Murugesan [a], David B.H. Tay [a,*], Ira Cooke [b], Pierre Faou [b]

[a] Department of Electronic Engineering, La Trobe University, Bundoora, Victoria 3086, Australia
[b] Department 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.

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

* Corresponding author.
E-mail addresses: s.murugesan@latrobe.edu.au (S. Murugesan), d.tay@latrobe.edu.au (D.B.H. Tay), i.cooke@latrobe.edu.au (I. Cooke), p.faou@latrobe.edu.au (P. Faou).

http://dx.doi.org/10.1016/j.compbiomed.2015.05.002
0010-4825/

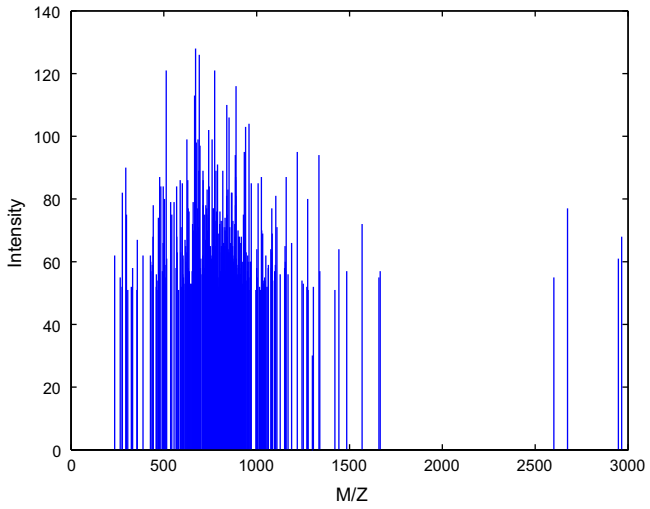

Fig. 1. Plot of a MS scan of the dataset 3666.

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 6. 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 isotope 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 nL/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 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