BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 18 2010, pages 2242–2249
doi:10.1093/bioinformatics/btq403

Sequence analysis

Advance Access publication July 13, 2010

A dynamic wavelet-based algorithm for pre-processing tandem mass spectrometry data

Penghao Wang[1,*], Pengyi Yang[2,3], Jonathan Arthur[4] and Jean Yee Hwa Yang[1]

[1]School of Mathematics and Statistics, [2]School of Information Technologies, University of Sydney, [3]National ICT Australia, Australian Technology Park and [4]Discipline of Medicine, Sydney Medical School, University of Sydney, Australia

Associate Editor: John Quackenbush

ABSTRACT

Motivation: Mass spectrometry (MS)-based proteomics is one of the most commonly used research techniques for identifying and characterizing proteins in biological and medical research. The identification of a protein is the critical first step in elucidating its biological function. Successful protein identification depends on various interrelated factors, including effective analysis of MS data generated in a proteomic experiment. This analysis comprises several stages, often combined in a pipeline or workflow. The first component of the analysis is known as spectra pre-processing. In this component, the raw data generated by the mass spectrometer is processed to eliminate noise and identify the mass-to-charge ratio (m/z) and intensity for the peaks in the spectrum corresponding to the presence of certain peptides or peptide fragments. Since all downstream analyses depend on the pre-processed data, effective pre-processing is critical to protein identification and characterization. There is a critical need for more robust pre-processing algorithms that perform well on tandem mass spectra under a variety of different conditions and can be easily integrated into sophisticated data analysis pipelines for practical wet-lab applications.

Result: We have developed a new pre-processing algorithm. Based on wavelet theory, our method uses a dynamic peak model to identify peaks. It is designed to be easily integrated into a complete proteomic analysis workflow. We compared the method with other available algorithms using a reference library of raw MS and tandem MS spectra with known protein composition information. Our pre-processing algorithm results in the identification of significantly more peptides and proteins in the downstream analysis for a given false discovery rate.

Availability: Software available at: http://www.maths.usyd.edu.au/u/penghao/index.html

Contact: penghao.wang@sydney.edu.au

Received on March 16, 2010; revised on June 24, 2010; accepted on July 04, 2010

1 INTRODUCTION

Mass spectrometry (MS)-based proteomics enables studies of the complexity and dynamics of proteins in biological systems (Anderson and Anderson 1998, Blackstock and Weir, 1999). It involves the identification and characterization of the entire set of proteins expressed by the genome as well as understanding changes in protein expression levels and how proteins interact with each other (Wilkins *et al.*, 1997). Increasingly, many scientific investigators are using tandem MS (MS/MS) for high-throughput protein identification and quantification. In recent years, there has been rapid development in MS technology. The development of methods, such as isobaric tag for relative and absolute quantitation (iTRAQ) (Unwin *et al.*, 2005), provides researchers with powerful techniques to determine the relative expression levels of thousands of proteins simultaneously. The precise identification of proteins is crucial in developing new diagnostic, prognostic and therapeutic products for the treatment of human diseases such as cancer, asthma and diabetes (Anderson and Anderson 2002). Furthermore, accurate quantitative protein expression information enables scientists to reliably determine the proteins critical to specific biological functions and provides a powerful mechanism to identify disease biomarkers and to design more effective medicines (Hanash, 2003). However, accurate statistical analysis of raw tandem spectrum data remains a challenging task.

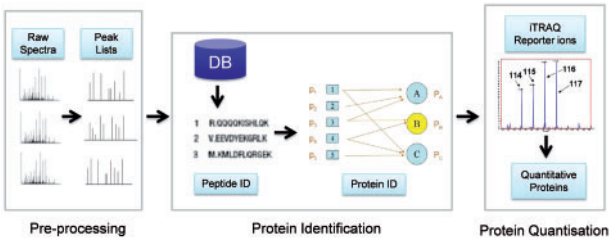

Fig. 1. A typical proteomic data analysis workflow and its major components. This includes pre-processing for peak determination, peptide and protein identification and quantization.

Current statistical and computational analysis of MS data involves several challenges. After acquiring raw spectra from a mass spectrometer, it is necessary to pass through all components of a complete data analysis workflow in order to obtain the final results. The complete data analysis workflow, as illustrated in Figure 1, can be sub-divided into several analysis components: spectra pre-processing, peptide and protein identification and protein quantification.

The first step in pre-processing involves identifying and locating peaks within raw MS and MS/MS spectra. These peaks (where by 'peaks' we refer to any peak-shaped signals) may correspond to the presence of peptides or peptide fragments in the sample.

*To whom correspondence should be addressed.

As the spectrum is usually tempered with a variety of other features, such as electrical and chemical noise, machine artifacts and contamination, the correct identification of peaks is difficult. Pre-processing must address each of these problems: this often involves several different procedures including: noise removal, baseline removal, peak detection, peak centroiding and intensity estimation. Depending on the MS ionization technology, peptides may also carry more than one charge. As this information is also critical to downstream protein identification algorithms, another important task for pre-processing algorithms is to correctly estimate the peptide charge state.

Pre-processing has great impact on the downstream protein identification and quantization analyses (Ong *et al*., 2003; Yu *et al*., 2005; Zhang *et al*., 2002). All protein database search engines and *de novo* sequencing algorithms (Cagney and Emili, 2002) depend on the quality of peak information used as input to their algorithms. Although there are a number of methods for pre-processing MS-level spectra, there is a lack of available algorithms specifically designed for MS/MS protein analysis workflow. Most available and widely used methods are based on signal intensity. These methods may produce unsatisfactory results and increase the false positive rate in real applications (Renard *et al*., 2009). Pre-processing should thus be properly addressed in order to obtain more reliable downstream protein identification and quantization analysis.

There are a number of available precursor-level MS pre-processing algorithms, and they may be broadly classified into three categories: (i) intensity-based approaches; (ii) peak modeling-based approaches; and (iii) wavelet-based methods.

Intensity-based approaches use certain thresholds to filter weak signals, leaving the most intense peaks. Such methods include mzWiff (Pedrioli *et al.*, 2004) provided by the Trans-Proteomic Pipeline (TPP) (Pedrioli, 2010), wiff2dta (Boehm *et al.*, 2004) and InsPecT (Tanner *et al.*, 2005). Many protein identification engines, for example OMSSA (Geer *et al.*, 2004) and X!Tandem (Craig and Beavis, 2004), apply a simple intensity-based method before initiating the protein identification search algorithm. Intensity-based methods can be improved by using predefined mass-to-charge ratio (m/z) intervals as in MaxQuant (Cox and Mann, 2008). These are the most commonly used methods and their main advantage is their simplicity. However, real peptide fragment signals can be weaker than spectrum noise. Thus, the limitation of these approaches is that they sometimes fail to detect peaks leading to a significant decrease in the performance of downstream analyses. Some methods apply signal filters before applying the intensity-based peak selection to improve accuracy. For example, MEND (Andreev *et al.*, 2003) uses a matched filter as the starting point (it also applies other techniques), PROcess (Li *et al.*, 2005) uses a moving average filter, MZmine (Katajamaa *et al.*, 2006) uses a Savitzky–Golay filter, and LIMPIC (Mantini *et al.*, 2007) uses a Kaiser Window Filter. However, these filters may cause distortion of the spectra and it is often hard to tell whether all the noise has been successfully filtered or whether a significant proportion remains in the signal.

Peak modeling-based approaches (Gentzel, 2003; Gras *et al.*, 1999; Lange *et al.*, 2006; Qu *et al.* 2003; Randolph and Yasui 2005) are pre-processing algorithms that try to take advantage of additional information other than signal intensity. As MS peaks have characteristic shapes and patterns based on the particular instruments used, the unique peptide signal shape provides a powerful means to identify real peaks from white and colored noise. These methods normally use peak width as the matching criterion to reduce false positives in peak identification. In real-life applications, this approach may become difficult because peaks have complex patterns, and various interferences such as noise can make the peak width estimation difficult. The peak shape and width depend on a number of factors. In addition, the width and height of real peaks can vary significantly across different spectra. Even in the same spectrum, peaks with high m/z values are usually wider and lower in amplitude compared to ones with low m/z values. Thus, simple and static peak models tend to produce highly variable and unreliable results in real applications (Du *et al.*, 2006). Some methods are based on models even involving more information. Gras *et al*. (1999) introduces an averaging model studying MALDI MS spectra; Gay *et al*. (1999) described a method using a theoretically peptide model. These approaches involve sophisticated algorithms and strong theoretical merits, and remain good options for pre-processing the MS-level spectra for which they were designed. Nevertheless, the fragmentation and peak patterns of MS/MS spectra are more complex than MS spectra and a more general approach is also desirable.

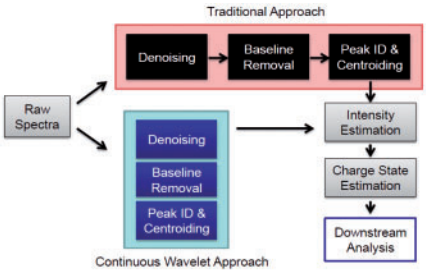

Fig. 2. Simplified procedure of CWT-based proteomic pre-processing.

Wavelet-based methods can be categorized into discrete wavelet (DWT) methods and continuous wavelet (CWT) methods. DWT methods are generally utilized as noise filters, e.g. Cromwell method (Coombes *et al.*, 2005). CWT methods can be utilized for several procedures of the pre-processing. First, it is possible to identify peaks without explicitly removing disturbing artifacts by using CWT analysis. With CWT methods, baseline, noise and the real signal can be separated by their different frequency ranges. A well-designed CWT method, which respects the specific characteristics of peptide signals, renders additional noise filters unnecessary, as noise-filtering is spontaneously achieved by wavelets. This is demonstrated in Figure 2. As shown, the procedure for a traditional pre-processing algorithm requires five separate steps. With CWT methods, the procedure can be simplified to three steps as given in Figure 2. Therefore, the peak shape and characteristics are easier to identify by wavelet coefficients. Another advantage of CWT methods is peak modeling. Du *et al.* (2006) described a method (MassSpecWavelet) which directly utilizes the CWT coefficients matrix of the spectrum to identify peaks. By using a range of scaled wavelets, the method can detect peaks with a lower false positive rate and better signal to noise ratio (SNR). The disadvantage of this method is the arbitrary selection of a large range of scales and the inability to select the most relevant scales. Such a static model may perform well in a specific situation; however, it may become difficult when wavelets are incorrectly selected and this may significantly increase false

positives. Thus it is desirable to have a good algorithm to correctly determine the best matching wavelets as the peak model.

Even though as mentioned there are a number of pre-processing methods available for MS-level data, many algorithms are designed for a specific type of instrument and do not support MS/MS spectra pre-processing. In addition, many methods were developed on different platforms, and are difficult to integrate into analysis pipelines. mzWiff is one of the few methods able to read the standard open format and support downstream proteomic analysis, although it is not unique in this regard. This lack of available downstream 'pipeline ready' MS/MS pre-processing algorithms seriously undermines the reliability of the protein identification and quantization.

We have developed a new wavelet-based MS and MS/MS pre-processing algorithm, called the Dynamic Wavelet Approach (DyWave), to address the shortcomings found in existing methods. It supports a wide array of instruments and it dynamically adjusts the peak model to achieve better performance. The algorithm is designed as an integrated component of the complete data analysis workflow (Wang *et al.*, 2009). Finally, our method does not detect the peaks using only intensity, but takes additional information regarding peak shape into account. In addition, it is one of the few algorithms that incorporate an effective method to estimate peptide charge.

In this article, we set out the statistical details of our model and its implementation. We then demonstrate the performance of our algorithm by comparing to other available methods using large-scale datasets obtained from different instruments. Our algorithm performs significantly better in the compared criteria including the number of correct identification at the peptide and protein level, the false discovery rates (FDRs) at both levels, and the final SNR. At the same FDR, up to 30% and 15% more proteins can be identified compared to methods provided in TPP and commercial software from Applied Biosystems.

2 MATERIALS AND METHODS

We introduce the dynamic peak model in our proposed method, followed by a brief step-by-step description of our method, and finally the evaluation study used to assess the performance of our method.

2.1 The dynamic wavelet peak model

Based on CWT, we propose a novel method that applies a dynamic model to achieve better accuracy. The CWT transform can be formulated as:

$$C(a,b)=\int_R s(t)\psi_{a,b}(t),\psi_{a,b}(t)=\frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right), \quad a\in R^+-\{0\}, b\in R, \tag{1}$$

where $s(t)$ is the signal, a is the scaling factor, b is the translation factor, $\psi_{a,b}(t)$ is the scaled and translated mother wavelet and C is the wavelet coefficient. Coefficients reflect the pattern matching between the signal s and the mother wavelet $\psi_{a,b}(t)$. The wavelet technique has an analytical advantage because it provides freedom in the choice of mother wavelets. By using different parameters, the daughter wavelets $\psi_{a,b}(t)$ can therefore provide a dynamic peak model without invoking more complicated non-linear curve fitting. For peak detection, the daughter wavelets should locally resemble the real signal. Gaussian family wavelets are very effective unless the peaks are strongly asymmetric. The real part of the Gaussian wavelet is

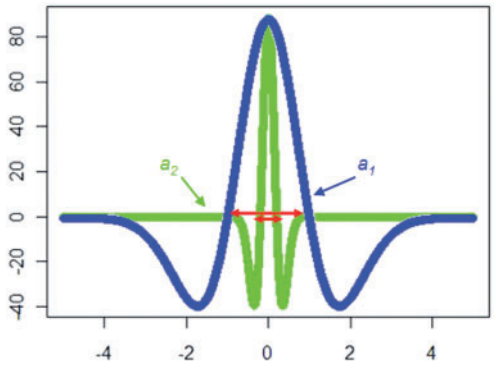

Fig. 3. The Marr wavelet under two different scaling factor a_1 and a_2.

calculated by taking the p-th derivative of the function:

$$\psi(t)=\frac{d^p\left(C_p e^{-t^2}\right)}{dt^p}, \tag{2}$$

where C_p is a constant factor which depends on the derivative order p of the function $C_p e^{-t^2}$ and is computed to normalize the Gaussian wavelet function. The Marr wavelet is proportional to the second derivative of the Gaussian wavelet function. It has been demonstrated (Lange *et al.*, 2006) that Gaussian wavelets are well suited for detecting individual peaks. If the scale is chosen correctly, the transform of a given peak is largely independent of neighboring peaks, even if they heavily overlap. The Marr wavelet can be formulated as:

$$\psi(t)=\frac{1}{\sqrt{2\pi a^3}}\left(1-\frac{t^2}{a^2}\right)e^{\frac{-t^2}{2a^2}}, \tag{3}$$

Figure 3 presents an example of Marr wavelets of scale a_1 and a_2, where the wavelet of a_2 has a smaller peak width, which is a better match for narrow peaks. Similarly, the wavelet of a_1, provides a better match for wider peaks. Our algorithm applies a linear function to describe the relationship between peak width and m/z value to achieve more reliable peak identification since the resolution of a mass spectrometer only depends on the instrument:

$$pw=Ax+B, \tag{4}$$

where pw is the peak width, x is the m/z, A and B are predefined constants based on the type of mass spectrometer used. Depending on the peak width and the spectrum m/z coverage, the peak model used is dynamically adjusted.

When the scaled wavelet resembles the peak, the coefficients will demonstrate a local maximum around the position of the peak centroid. The coefficient becomes stronger when the scaled wavelet more closely resembles the peak. Figure 4A demonstrates a single MS/MS peak with the centroid marked by a red cross. Figure 4B is the distribution of the amplitude of the coefficient maxima across different scales. It shows the strongest coefficient maximum happens at the scale the daughter wavelet best matches the peak. When the scale becomes smaller or larger, the coefficient gradually reduces in amplitude. The coefficient maximum approximately follows a Gaussian process in the continuous space as shown. To increase reliability, our algorithm transforms the spectra on the wavelets that best fit the peaks. If the peak width at the smallest m/z is estimated to be w_1 while for the largest m/z it is estimated to be w_2, then the best wavelet scales will be within the interval $[0.5\times w_1, 2\times w_2]$. Based on the predefined number of allowed scales N, the scaling interval can be calculated by $N(2\times w_2-0.5\times w_1)$. In order to further increase accuracy, our algorithm incorporates a Gaussian weighting model to select the best matched daughter wavelets, because the performance of the wavelet-based peak modeling largely depends on the correct selection of the daughter wavelets. For a specific peak at a certain m/z, the method selects the wavelets of the scales from half to double the

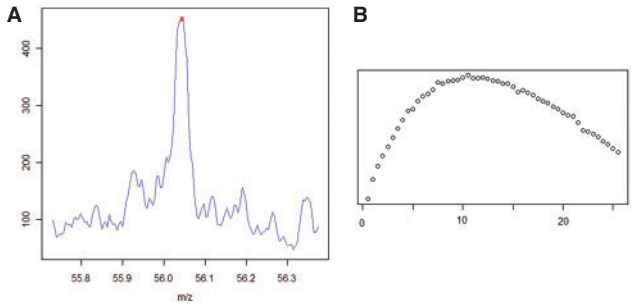

Fig. 4. (**A**) A single MS/MS peak with the centroid marked by a red cross. (**B**) The distribution of the amplitude of corresponding wavelet coefficient maximum.

peak width $[0.5 \times pw, 2 \times pw]$. The Gaussian weight can be calculated by:

$$w_i = (-1)^p \frac{1}{\sqrt{2\pi}} e^{\left(0.5\left(\frac{2^*(s_i - s_1)}{(N-1)\mathrm{SI}} - 1\right)^2\right)}, \quad (5)$$

where w_i is the weight for a specific coefficient maximum at scale i, and s_i and s_1 are the scale value of i and the smallest considered scale respectively. N is the total number of considered scales and SI is the scale interval. P is a penalty factor: if there is no relevant coefficient maximum present at that scale, P is 1; if the maximum is present, P is 0. When the accumulated weight of all the coefficient maxima of all the chosen scales is larger than a threshold, then a peak candidate is detected.

2.2 DyWave peak identification procedure

Our dynamic wavelet-based pre-processing algorithm (DyWave) involves five steps and Box 1 provides an outline.

Step 1: Extract the raw MS/MS and precursor MS spectra from standard mzXML files. Estimate the smallest interval of the spectra and thus obtain the best matching wavelet using the linear function as Equations (3) and (4).

Step 2: Identify the local maxima of the coefficients using a sliding window on the regions of interest. The size of the sliding window is linearly determined by the wavelet scale. Then calculate the Gaussian weights of all identified coefficient maxima using Equation (5). If the final weight is larger than a defined threshold, it is considered a candidate peak.

Step 3: Estimate the peak centroids in the coefficient domain instead of intensity, because the baseline and noise have been suppressed or removed. Two options are provided. One way is to use the averaged position as the peak centroid. The other is to apply the same Gaussian weighting scheme:

$$C = \frac{(c_1 w_1 + c_2 w_2 + \ldots + c_n w_n)}{N}, \quad (6)$$

where c_i is the coefficient maximum position at scale i, w_i is the associated weight, and N is the total number of considered scales. This may provide a better approximation of the peak centroid.

Step 4: Refine the peak list using two criteria: (i) the selected peaks should have an estimated SNR larger than a predefined threshold, default is 3, and/or (ii) the distance between two adjacent peak positions should also exceed a threshold. Then estimate the peak SNR by using the wavelet coefficients. For a specific peak, the signal is defined as the coefficient maximum amplitude at the best matching scale. The coefficients at the scale where the smallest calculated wavelet, and the 95% quantile of the coefficient intensity around the considered peak is calculated as noise (Du *et al.*, 2006).

Step 5: Estimate peak intensity using two possible ways. First, the intensity at the determined peak centroid is used. The second approach is to use the area under the curve (AUC) as the intensity instead of the value of one data point. To calculate the AUC, DyWave applies a moving average method to find the two approximate end positions of the peak and then calculates the AUC for that specific peak using the signal intensity. Our method incorporates the additional (optional) feature of estimating the peptide charge state by analyzing the MS and MS/MS spectra. This feature can facilitate the downstream analysis especially on ESI datasets.

Box 1. The dynamic wavelet-based pre-processing algorithm

Input: standard mzXML spectra files
Output: peak lists in identification softwares compatible formats
for each MS/MS spectrum **do**
  obtain optimal peak width range by Equation (4)
  obtain optimal wavelet scales from $\mathrm{scale}_{\min}$ to $scale_{\max}$, total of N scales
  calculate wavelet coefficients Coef_1 on $\mathrm{scale}_{\min}$ **do**
  for each sliding window on spectrum of $\mathrm{scale}_{\min}$
    find local coefficient maxima and initilize as peak list
  endfor
  for each scale i from $\mathrm{scale}_{\min}+1$ to $\mathrm{scale}_{\max}$ **do**
    calculate wavelet coefficients $\mathrm{Coef}_{i,j}$
    find local coefficient maxima $\mathrm{LocalMax}_{i,j}$
    append maxima $\mathrm{LocalMax}_{i,j}$ to peak list $\mathrm{Peak}_{i,j}$
    calculate Gaussian weights $\mathrm{Weight}_{i,j}$ by Equation (5)
  end for
  calculate the total weight for candidate peaks $\mathrm{totalWeight}_i = \sum_i \mathrm{Weight}_{i,j}$
  if $\mathrm{totalWeight}_i <$ threshold **do**
    remove element from peak list
  end if
  estimate SNR for each peak
  if SNR < threshold **do**
    remove this element from peaks
  end if
  if peak is too close to neighbor peak **do**
    merge this peak
  end if
  estimate peak centroids by average or Equation (6)
  for each refined peak in Peak_i **do**
    find two ends of the peak
    calculate the AUC as peak intensity
  end for
end for

2.3 Evaluation

To evaluate the performance of different pre-processing algorithms, we examine the effect of the different algorithms on downstream protein identification results. We use the raw spectra from two large-scale datasets as a benchmark: (i) the Aurum dataset (Falkner *et al.*, 2007), and (ii) human protein mixture study datasets from the Clinical Proteomic Technologies Assessment for Cancer (CPTAC) (http://cptac.tranche.proteomecommons.org). The Aurum dataset is a public, open library of MS and MS/MS spectra generated on an ABI 4700 MALDI TOF/TOF from known purified and trypsin digested protein samples. The acquisition procedure utilizes a workflow used for gel-purified proteins. To our knowledge, the Aurum dataset is one of few large, publicly available MS and MS/MS reference datasets where the raw spectra are provided and the actual identity of the proteins is known in advance of the analysis. The CPTAC dataset comes from a large-scale study of the reproducibility and repeatability of National Cancer Institute (NCI) and Universal Proteomics Standard set 1 (UPS1) human proteins. The UPS1 comprises of 48 known human proteins and NCI dataset comprises of 20 known human proteins. These proteins were analyzed in five different concentrations and in mixture. We also evaluate the performance of our method on these datasets which were obtained using a LTQ-Orbitrap from ThermoFinnigan. The MS and MS/MS spectra of NCI-20 samples and UPS1 samples were combined for pre-processing and protein identification. The details of CPTAC datasets can be found in Tabb *et al.* (2010).

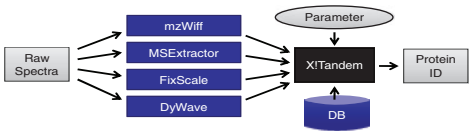

Fig. 5. The design of the evaluation experiment: all parameters are controlled in order to evaluate pre-processing algorithms.

We compare DyWave with three other widely used algorithms:

(1) The intensity-based approach mzWiff from the TPP. This algorithm was chosen because, to our knowledge, it is the only open source MS/MS pre-processing algorithm. In addition, it is possibly the most widely used method for spectrum pre-processing.

(2) The commercial software MSExtractor from Applied Biosystems for the ABI4700 MALDI TOF/TOF mass spectrometer (Falkner *et al.*, 2007). Proprietary software normally involves sophisticated algorithm design optimized for the supporting instrument. Thus it should produce very reliable results. MSExtractor can be only used on the Aurum dataset because it is specific to datasets from this manufacturer.

(3) Our own implementation (FixScale) of a CWT approach similar to the method of Du *et al.* (2006) using a static wavelet scale from 1 to 64 and detecting peaks by linking the maxima across the scale coefficients which are then filtered by the condition: SNR>3.

Once peak lists are obtained using these algorithms, X!Tandem (Craig and Beavis, 2004) is used to perform protein identification. The default search parameters are used and searches are conducted against the SWISS-PROT human database. The experimental design is presented in Figure 5. The performance of the pre-processing algorithms is evaluated by comparing the peptide and protein identification results.

There are three potential benefits to be derived from improved pre-processing: (i) more peptides and thus more proteins could be identified; (ii) the FDR in protein identification could be reduced; and (iii) spectra SNR could be increased and noise reduced. Therefore, these three criteria are used to evaluate the various methods.

A key aim of proteomic analysis is to identify the maximum possible number of peptides and proteins while controlling the FDR. The FDR can be estimated in two ways. First, as the proteins in the Aurum dataset and CPTAC datasets are known in advance, the FDR can be calculated directly. However, this approach is only feasible with reference datasets of known composition. Thus, we also use a second target-decoy approach to estimate the FDR. Briefly, the method constructs a decoy database by reversing all protein sequences in the original database and concatenating this set of reversed sequences to the original database. This combined database is used in the protein identification search. An estimate of the FDR is then obtained by doubling the number of decoy hits and dividing by the number of total hits: FDR= 2× DecoyHits/TotalHits (Elias and Gygi, 2007).

3 RESULT

The effectiveness of the various pre-processing algorithms is demonstrated by examining the peptide-level results, the protein-level results and the SNR results in turn.

3.1 Peptide-level results

Peptide identification is directly associated with spectra quality. Thus, the results at the peptide level are a more direct indication of the pre-processing performance with less bias. At the peptide level, DyWave performs significantly better compared to the other methods. This is shown in Figure 6. Using the Aurum dataset,

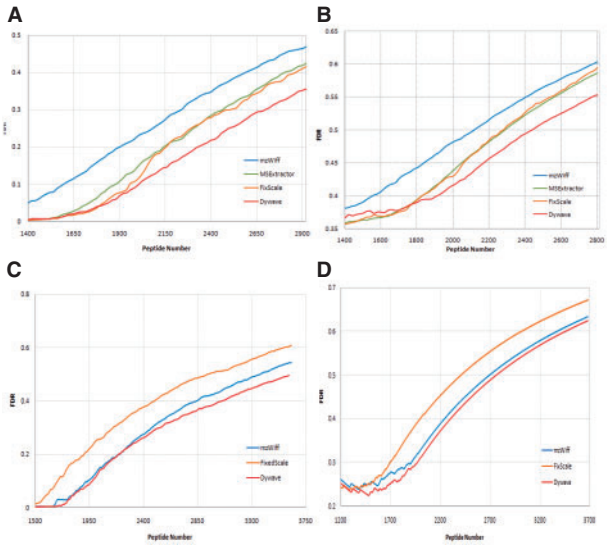

Fig. 6. The peptide identification results. (**A**) FDR estimated by the target-decoy approach on the Aurum dataset. (**B**) Real FDR directly calculated on the Aurum dataset. (**C**) FDR estimated by the target-decoy approach on the Orbitrap dataset. (**D**) Real FDR directly calculated on the Orbitrap dataset.

DyWave identifies approximately 15% more unique peptides for a given false positive rate compared to the proprietary software and up to 30% more peptides compared to the commonly used intensity-based approach. The performance of DyWave is consistently better than other compared methods on the CPTAC datasets and accordingly identifies significantly more peptides. This indicates that the processed spectra obtained by DyWave are of much higher quality, greatly facilitating the successful identification of the peptides from the tandem MS spectra. As shown in Figure 6, DyWave has the highest sensitivity, and it is able to achieve better accuracy than the static algorithm based on a fixed scale wavelet transform.

Using the target-decoy approach, the estimated peptide FDR of mzWiff is much higher than that of the other two methods, especially on Aurum dataset. As shown in Figure 6A, the commercial algorithm, fixed scale method and DyWave can achieve almost 100% specificity when identifying 1400 or less peptides. On the other hand, the intensity-based method can never reach such accuracy even when identifying 100 peptides. This is consistent with the results using the Aurum dataset that show the other three methods perform much better than mzWiff as shown in Figure 6B. On the LTQ-Orbitrap datasets, DyWave consistently performs better than the other methods, including the fixed scale wavelet method and mzWiff. Comparing Figure 6C to A shows the intensity-based approach performs better on the LTQ-Orbitrap data though still incurring higher FDR than DyWave. These results clearly demonstrate pre-processing has a great impact on the downstream peptide identification analysis. It also indicates using signal intensity alone may not always provide satisfactory results and is prone to incur a higher false positive rate.

Comparing the target-decoy FDR results in Figure 6A and C to the directly calculated FDR results in Figure 6B and D, it clearly demonstrates that the target-decoy strategy tends to underestimate the actual false positive rates. This phenomenon is consistent on both

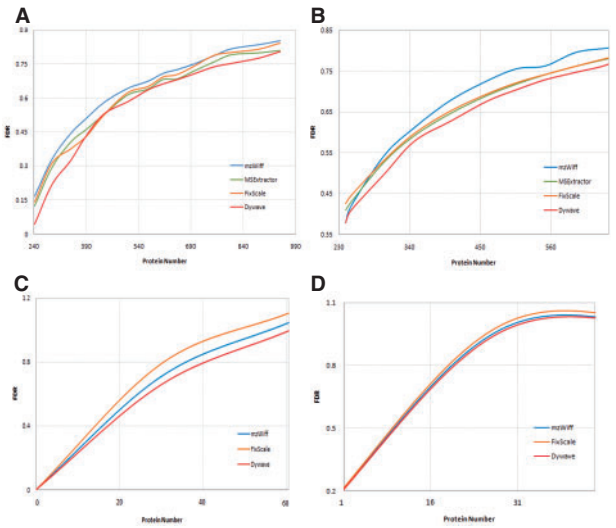

Fig. 7. The protein level results. (**A**) FDR estimated by target-decoy approach on the Aurum dataset. (**B**) Real FDR directly calculated on the Aurum dataset. (**C**) FDR estimated by the target-decoy approach on the Orbitrap dataset. (**D**) Real FDR directly calculated on the Orbitrap dataset.

MALDI TOF/TOF and LTQ-Orbitrap instruments. This indicates there is unequal likelihood of selecting an incorrect peptide match from the target database compared to the reversed decoy database. In other words, false positive target hits are more likely than decoy hits. On the other hand, the distribution of the real FDR and target-decoy FDR is consistent as shown in Figure 6. This indicates that the target-decoy strategy is able to capture the random matched hits contribution of the false positive identification.

3.2 Protein-level results

The protein-level results are consistent with the peptide level results. DyWave performs significantly better than the three compared methods. In general, DyWave is able to identify ∼10% more true unique proteins without incurring a higher false positive rate. For each identification at the protein level, DyWave achieves higher confidence compared to the other methods. As DyWave identifies a larger number of proteins for the same FDR at the peptide level, each identification at the protein level receives, on average, more peptide identification support. As Figure 7 demonstrates, on the Aurum dataset the intensity-based mzWiff again identifies the smallest number of proteins at a given FDR, while DyWave identifies the largest number of proteins at the same FDR and the proprietary software and fixed-scale wavelet method sit in between. On the CPTAC datasets, DyWave consistently identified the most proteins while the fixed-scale wavelet method identified the smallest number and mzWiff is in the middle.

Based on the peptide- and protein-level results, DyWave performs consistently better than the mzWiff method offered by TPP. It is interesting to note that, while the proprietary algorithm identifies more peptides than mzWiff on the Aurum dataset, this fails to translate to more protein identifications. The difference between the intensity approach, fixed scale wavelet method and the commercial software is marginal at the protein level and the performance of these three methods becomes even closer as the FDR increases

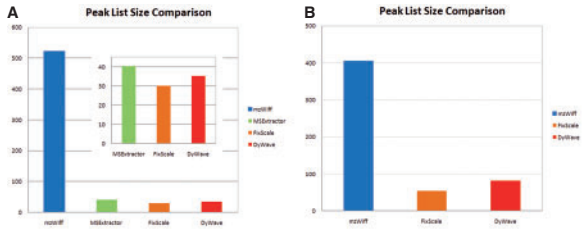

Fig. 8. The processed spectra file size comparison, where DyWave is shown to achieve the best SNR. (**A**) Comparison on the Aurum dataset. (**B**) Comparison on the LTQ-Orbitrap dataset.

as shown in Figure 7A and B. On the CPTAC datasets, mzWiff produces better results compared to its results on the Aurum dataset as Figure 7C and D show. On the other hand, DyWave presents the best performance on all datasets. This may be due to many peaks corresponding to real *y*- or *b*- ions having relatively weak amplitude and thus being difficult to identify. It is interesting to see that our dynamic wavelet method performs better than the static algorithm using a fixed wavelet scale. This demonstrates that the dynamic model with a more realistic weighting scheme can better capture and recognize the characteristics of the peaks. This is a strong indication that DyWave not only improves the quality of the processed spectra, it also can increase the number of actual peaks in the spectrum correctly identified, known as spectra identification coverage. In addition to increasing the number of proteins identified from a sample, this feature may also improve quantitative analysis, since quantitative technologies such as iTRAQ are entirely dependent on the correct identification of reporter ions.

Using the target-decoy approach, DyWave is able to identify up to 20% more unique proteins compared to mzWiff and 10% more proteins than MSExtractor and the fixed scale wavelet method at the same protein FDR.

Protein-level results demonstrate the target-decoy method provides a better estimate of the FDR than it does at peptide level. Comparing the estimated protein FDR and the actual FDR in Figure 7, we can see that the target-decoy estimated FDR matches the actual distribution of the FDR fairly well. However, it again demonstrates that the target-decoy approach is prone to underestimate the FDR even though the estimate seems to be more realistic at the protein level. When we examine the difference of the FDR at peptide and protein level, it is obvious that although peptide identification determines the quality of the downstream protein level analysis, the sensitivity of protein identification displays a different distribution. It is worth noting that while the FDR at peptide level remains constantly low and starts to grow exponentially after a certain stage, the protein-level FDR seems to follow a logarithmic distribution.

3.3 SNR and noise reduction

The SNR of pre-processed spectra can be estimated by comparing the file size of the peak lists giving rise to similar numbers of identified proteins. mzWiff can reduce the information in the raw spectrum by 60–90%. As Figure 8 demonstrates, this can be reduced by up to another 90% by the commercial software and the two wavelet-based methods. This significant reduction of file size is achieved without compromising the useful information. This indicates that the proprietary algorithm and wavelet methods achieve

much better SNR compared to the intensity-based mzWiff. The two wavelet methods can reduce the peak list by a further 10% compared to the commercial software while achieving even better downstream peptide and protein identification results. Even though the fixed-scale wavelet method is able to achieve the best results in reducing the file size, it also reduces the number of identified peptides and proteins. This indicates that real peptide fragment signals may be aggressively removed in the fixed-scale wavelet method. This demonstrates that DyWave achieves better SNR than the fixed-scale wavelet method without sacrificing more useful information. This may greatly improve the reliability of downstream analysis.

A better SNR has two additional benefits. First, because MS experiments generate huge volumes of spectra data, peak lists with higher SNRs require less storage space and make spectra transfer more convenient. Second, higher SNRs and smaller file sizes result in a significant increase in the computational speed of downstream analysis. This will greatly facilitate the high-throughput proteomic analysis applications and automatic analysis pipelines.

4 DISCUSSION

In this article, we examined the overall proteomic data analysis workflow and demonstrated that our newly proposed dynamic wavelet-based pre-processing algorithm is able to significantly outperform currently available MS/MS pre-processing algorithms by producing more accurate peptide and protein identifications and increasing the spectra SNR. These are the key functionalities for a pre-processing algorithm. Our experiments demonstrate that mass spectra pre-processing is an important component of the overall MS-based proteomic analysis, greatly affecting the downstream results. Considerably more peptides and proteins can be identified with more sophisticated pre-processing algorithms. With powerful spectra pre-processing methods, the accuracy and reliability of MS-based quantitative proteomic analysis, such as isotope-coded affinity tag (ICAT), stable isotope labeling with amino acids in cell culture (SILIAC) and iTRAQ can also be greatly improved. Despite this, pre-processing has attracted less attention than the other aspects of tandem MS data analysis and there is a lack of available, reliable pre-processing algorithms for tandem MS data. Therefore, we have developed our new pre-processing algorithm to address this issue.

Spectrum signal intensity has been the major criterion used in pre-processing (peak picking) tandem MS spectra. Consequently, peptide and protein identification has been critically dependent on the spectrum signal intensity. The superior performance of DyWave demonstrates that such an approach has its limitations. We have shown that using additional information about signal shape can lead to much more accurate identification and increase identification coverage. This is especially true with low precision mass spectrometers where excessive noise and spurious peaks are expected. Even with high precision instruments, it is common for many peaks corresponding to real *y*- or *b*- peptide fragment ions to have relatively weak intensity, thus making them difficult to identify. This may be why our method is able to identify more peptides and proteins by taking account of additional signal shape characteristics. Assessment results also demonstrate that accurate selection of the wavelet transform scale is crucial. With more accurately selected wavelet transform scales; a dynamic wavelet model can have better performance than a static wavelet method which applies wavelet transform on fixed scales. Therefore, the applicability of intensity-based pre-processing becomes difficult in many real applications, and methods having stronger analytical basis, such as wavelet-based approaches, are more reliable.

High-throughput tandem MS proteomic analysis is peptide-centric: the identification and quantification of proteins are inferred from the identified peptides giving rise to the observed spectra. In most cases, peptide-level analysis is an intermediate step because the ultimate goal of most experiments is to identify the proteins. However, the ease of protein identification often depends on the number and quality of the identified peptides. In comparing protein-level identification and peptide-level identification, the commercial software cannot consistently translate an improvement in peptide identification to an improvement in protein identification although it can identify more peptides than the predominant intensity-based approach from the TPP. Our method shows a clear advantage from this perspective. This may be due to larger number of weak peptide fragment signals being identified by our method. This is hard to attain using the static wavelet method. Furthermore, these weak peptide signals enable the protein database search engine to successfully assign more protein identifications. Thus, we hypothesize that the successful identification of proteins is dependent on 'two' distinct aspects: the identification of peptides from spectra and the peptide to protein assignment. One potential reason for the independence of these two aspects is ambiguities in determining the identities of proteins that share multiple peptides. Thus identifying more peptides is not the only benefit we can gain from a dynamic pre-processing algorithm. By successfully recovering the weak peptide fragments from the spectra, we can greatly improve the peptide to protein assignment.

As protein identification can be an error-prone exercise, the estimation of the frequency of false identification is important (Keller *et al*., 2002). The target-decoy evaluation strategy has become the most widely used means of estimating the FDR in proteomic research. This approach involves introducing answers that are known *a priori* to be incorrect, called 'decoys', to the search space. By making the assumption that incorrect identifications are uniformly distributed in the search space, one can estimate the FDR from the number of decoy hits. Existing study has indicated that the target-decoy strategy is prone to underestimate the actual FDR, especially at peptide level (Käll *et al*., 2008). The underestimated FDR suggests the distribution of decoy identifications does not accurately represent the target 'null' distribution. In other words, there is unequal likelihood of selecting an incorrect peptide match from the target database as compared to selecting a match from the reversed decoy database, even if the search algorithm is presented with an equal number of target and decoy peptides. In our comparison studies, we have demonstrated the estimated FDRs in a comparative setting are biased toward underestimating the real FDRs in the identification, even though a reversed sequence database seemingly is the logical choice for a decoy. False positive target hits have been demonstrated to be more likely than decoy hits. Therefore, we should keep in mind that the real FDR is likely to be higher than the target-decoy estimated FDR.

One difficulty in proteomic research is most software is proprietary and there is a lack of open alternatives. This problem significantly hinders the advance of proteomic research as proprietary software often makes collaborative research difficult. Many modern bioinformatics research projects require the joint

effort of many individual laboratories, e.g. the human genome project. To this end, it is necessary to have open source analysis applications that can be shared and improved by all participants. Furthermore, although proprietary software is typically reliable, it is developed based on the knowledge and expertise of one team or one company. The inability to examine the specific details of the underlying algorithm means the software is not able to be improved or developed using ideas from the hundreds of outside experts.

In conclusion, pre-processing is an important component of the proteomic analysis workflow. It has a great influence on the success of downstream analysis components. More advanced pre-processing algorithms are desirable since they result in considerably more peptide and protein identification and a higher SNR. Dynamic models using wavelet theory provides a powerful means for pre-processing raw spectra. The DyWave method will be freely available for academic purposes.

ACKNOWLEDGEMENTS

Our method is implemented in C++. The implementation adopts source code from TPP for processing mzXML data format. We thank Dr Ben Crossett, Philippa Kohnke and Prof. Richard Christopherson from University of Sydney for providing discussion and initial data during the course of the work. We thank Mark Gjukich and Dr Phillip Andrews from University of Michigan for providing Aurum dataset and raw spectra.

Funding: ARC Discovery Grant (DP0984267); NICTA scholarship (to P.Y.).

Conflict of Interest: none declared.

REFERENCES

Anderson,N.L. and Anderson,N.G. (1998) Proteome and proteomics: new technologies, new concepts, and new words. *Electrophoresis*, **19**, 1853–1861.

Anderson,N.L. and Anderson,N.G. (2002) The human plasma proteome: history, character, and diagnostic prospects. *Mol. Cell. Proteom.*, **1**, 845–867.

Andreev,V. *et al*. (2003) A new algorithm for minimizing chemical noise in LC-MS: matched filtration with experimental noise determination (MEND). In *Proceedings of the 51st ASMS Conference on Mass Spectrometry and Applied Topics*. Montreal, Quebec, Canada.

Blackstock,W.P. and Weir,M.P. (1999) Proteomics: quantitative and physical mapping of cellular proteins. *Trends Biotechnol.*, **17**, 121–127.

Boehm,A.M. *et al*. (2004) Extractor for ESI quadrupole TOF tandem MS data enabled for high throughput batch processing. *BMC Bioinformatics*, **5**, 162.

Cagney,G. and Emili,A. (2002) De novo peptide sequencing and quantitative profiling of complex protein mixtures using mass-coded abundance tagging. *Nat. Biotechnol.*, **20**, 163–170.

Coombes,K.R. *et al*. (2005) Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform. *Proteomics*, **5**, 4107–4117.

Cox,J. and Mann,M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.*, **26**, 1367–1372.

Craig,R. and Beavis,R. (2004) TANDEM: matching proteins with mass spectra. *Bioinformatics*, **20**, 1466–1467.

Du,P. *et al*. (2006) Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics*, **22**, 2059–2065.

Elias,J. and Gygi,S. (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods*, **4**, 207–214.

Falkner,J.A. *et al*. (2007) Validated MALDITOF/TOF mass spectra for protein standards. *J. Am. Soc. Mass Spectr.*, **18**, 850–855.

Gay,S. *et al*. (1999) Modeling peptide mass fingerprinting data using the atomic composition of peptides. *Electrophoresis*, **20**, 3527–3534.

Geer,L.Y. *et al*. (2004) Open mass spectrometry search algorithm. *J. Proteome Res.*, **3**, 958–964.

Gentzel,M. *et al*. (2003) Preprocessing of tandem mass spectrometric data to support automatic protein identification. *Proteomics*, **3**, 1597–1610.

Gras,R. *et al*. (1999) Improving protein identification from peptide mass fingerprinting through a parameterized multi-level scoring algorithm and an optimised peak detection. *Electrophoresis*, **20**, 3535–3550.

Hanash,S. (2003) Disease proteomics. *Nature*, **422**, 226–232.

Katajamaa,M. *et al*. (2006) MZmine: Toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics*, **22**, 634–636.

Keller,A. *et al*. (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem*., **74**, 5383–5392.

Käll,L. *et al*. (2008) Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J. Proteome Res.*, **7**, 29–34.

Lange,E. *et al*. (2006) High-accuracy peak picking of proteomics data using wavelet techniques. In *Proceedings of Pacific Symposium on Biocomputing*, Maui, Hawaii, USA, pp. 243–254.

Li,X. *et al*. (2005) SELDI-TOF mass spectrometry protein data. In Gentleman,R. *et al*. (eds) *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer, New York, USA, pp. 91–109.

Mantini,D. *et al*. (2007) LIMPIC: a computational method for the separation of protein MALDITOF-MS signals from noise. *BMC Bioinformatics*, **8**, 101.

Ong,S.E. *et al*. (2003) Mass spectrometric-based approaches in quantitative proteomics. *Methods*, **2**, 124–130.

Pedrioli,P. *et al*. (2004) A common open representation of mass spectrometry data and its application to proteomics research. *Nat. Biotechnol.*, **22**, 1459–1466.

Pedrioli,P. (2010) Trans-proteomic pipeline: a pipeline for proteomic analysis. *Methods Mol. Biol.*, **604**, 213–238.

Qu,Y. *et al*. (2003) Data reduction using a discrete wavelet transform in discriminant analysis of very high dimensionality data. *Biometrics*, **59**, 143–151.

Randolph,T.W. and Yasui,Y. (2006) Multiscale processing of mass spectrometry data. *Biometrics*, **62**, 589–597.

Renard, B.Y. *et al*. (2009) When less can yield more - Computational preprocessing of MS/MS spectra for peptide identification. *Proteomics*, **9**, 4978–4984.

Tabb,D.L. *et al*. (2010) Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. *J. Proteome Res*., **9**, 761–776.

Tanner,S. *et al*. (2005) InsPecT: identification of posttransiationally modified peptides from tandem mass spectra. *Anal. Chem*., **77**, 4626–4639.

Unwin,R.D. *et al*. (2005) Quantitative proteomic analysis using isobaric protein tags enables rapid comparison of changes in transcript and protein levels in transformed cells. *Mol. Cell. Proteom.*, **4**, 924–935.

Wang,P. *et al*. (2009) An integrative approach to iTRAQ analysis. In *Proceedings of Bioinformatics*. Australia, Melbourne.

Wilkins,M. *et al*. (1997) *Proteome Research: New Frontiers in Functional Genomics*. 1st edn. Springer, Berlin, Germany.

Yu,W. *et al*. (2005) Statistical methods in proteomics. In Pham,H. ed. *Springer Handbook of Engineering Statistics*. Springer, London, UK, pp. 623–638.

Zhang,N. *et al*. (2002) ProbID: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. *Proteomics*, **2**, 1406–1412.