**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 proteomics 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 peaks within raw MS and MS/MS spectra. These peaks (where the ‘peaks’ refer to any peak-shaped signals) may correspond to the presence of peptides or peptide fragments in the sample. 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.

**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.

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.
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 mZiff (Pedioli et al., 2004) provided by the Trans-Proteomic Pipeline (TPP) (Pedioli, 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, MEzine (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.

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
We introduce the dynamic peak model in our proposed method, followed by using different parameters, the daughter wavelets advantage because it provides freedom in the choice of mother wavelets. To provide a dynamic peak model without invoking more complicated non-linear curve fitting, 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.

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 from Applied Biosystems. 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.

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:

\[
C(a, b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} f(t) \psi_{a,b}(t) dt,
\]

where \( f(t) \) 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 \( f(t) \) and the mother wavelet \( \psi_{a,b}(t) \). The wavelet technique has an analytical coefficient. The Marr wavelet can be defined as:

\[
\psi(t) = \frac{\partial^p C^p e^{-t^2}}{\partial t^p},
\]

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:

\[
\psi(t) = \frac{\partial^3 C^3 e^{-t^2}}{\partial t^3}.
\]

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:

\[
pw = Ax + B,
\]

where \( pw \) is the peak width, \( x \) is the \( m_z / C \), and \( A \) 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 peak centroid is estimated to be \( w_2 \), while for the largest scale \( w_1 \); it is estimated to be \( w_1 \). 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 scale, the method selects the wavelets of the scales from half to double the

\[
\frac{m_z}{C} = \frac{1}{2} w_1, \frac{3}{2} w_1, \ldots, \frac{9}{2} w_1
\]

...
Two options are provided. One way is to use the averaged position as the peak width \( 0 \times pw, 2 \times pw \). The Gaussian weight can be calculated by:

\[
w_i = \left( \frac{1}{\sqrt{2\pi}} \right) e^{-\frac{(x - x_i)^2}{2\sigma_i^2}}
\]

where \( w_i \) is the weight for a specific coefficient maximum at scale \( i \), and \( x_i \) and \( \sigma_i \) 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 at the determined peak centroid. The other is to apply the same Gaussian weighting scheme:

\[
C_i = \frac{1}{N} \sum_i w_i
\]

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 amplitude is used 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 scale_{min} to scale_{max}, total of \( N \) scales
- calculate wavelet coefficients Coef_{i,j} on scale_{i,j}
- for each sliding window on spectrum of scale_{i,j} do
  - find local coefficient maxima and initialize as peak list
  - for each scale \( i \) from scale_{min} + 1 to scale_{max} do
    - calculate wavelet coefficients Coef_{i,j}
    - find local coefficient maxima LocalMax_{i,j}
    - append maxima LocalMax_{i,j} to peak list Peak_{i,j}
    - calculate Gaussian weights Weight_{i,j} by Equation (5)
  - end for
  - calculate the total weight for candidate peaks totalWeight = \( \sum_i \) Weight_{i,j}
  - remove element from peak list
  - 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,j} 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 tryptic 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 Standards 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 UPS1-20 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).
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, 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 datasets.

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 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 to the original database. This combined database is used in the peptide and protein identification search. An estimate of the FDR is then obtained by comparing this approach is only feasible with reference datasets of known composition.

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 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).
The protein-level results are consistent with the peptide level results. The performance of the proprietary algorithm and wavelet methods achieve a larger number of proteins for the same FDR at the protein level. Each identification at the protein level is more confident compared to the other methods. As DyWave identifies the smallest number of proteins at the same FDR, 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 fixed-scale wavelet method sit in between. On the CPTAC datasets, mZWiff is in the middle. 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 identifies 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.
We have shown that using additional information about signal performance than a static wavelet method which applies wavelet transform scales; a dynamic wavelet model can have better wavelet transform scale is crucial. With more accurately selected proteins by taking account of additional signal shape characteristics. have relatively weak intensity, thus making them difficult to identify. many peaks corresponding to real mass spectrometers where excessive noise and spurious peaks are identification coverage. This is especially true with low precision DyWave demonstrates that such an approach has its limitations. 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 TTP. 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 TFP 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 Gjikich and Dr Philip 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