A New Hybrid De Novo Sequencing Method For Protein Identification

Penghao Wang[1*], Albert Zomaya[2], Susan Wilson[1,3]

1. Prince of Wales Clinical School, University of New South Wales, Kensington NSW 2052, Australia
2. School of Information Technologies, University of Sydney, Camperdown NSW 2006, Australia
3. Mathematical Sciences Institute, Australian National University, Canberra ACT 0020, Australia
*. Corresponding author
Email: penghao.wang@unsw.edu.au; albert.zomaya@sydney.edu.au; sue.wilson@anu.edu.au

Abstract—**Tandem mass spectrometry is a powerful tool for studying proteins. However, an open problem for proteomics research is how to accurately identify proteins from the experimental mass spectra. De novo sequencing based protein identification is the only feasible approach for finding new proteins and studying protein post-translational modifications. In this paper, we describe our novel hybrid de novo sequencing based protein identification method. It differs from existing methods which rely on finding one maximum path from a spectrum graph. Instead, to identify peptides, our method applies a novel Bayesian network and dynamic programming hybrid algorithm to explore the sub-optimal space. Thus our method can better accommodate various interferences and artefacts present in the mass spectra. Evaluated on a large number of spectra, our method outperforms the most popular de novo sequencing methods and can significantly improve the accuracy of de novo sequencing based protein identification.**

Keywords-Protein identification, de novo sequencing, Bayesian network, dynamic programming, proteomics.

I. INTRODUCTION

In recent years, tandem mass spectrometry (MS/MS) has become the leading technology for proteomics research [1, 2]. In a single mass spectrometry (MS) experiment, thousands of proteins from multiple complex biological samples can be identified and their expressions accurately measured at nano-mol level, thus providing a high throughput and high sensitivity approach for proteomics research. In a typical MS experiment, samples are first mixed and treated with proteolytic enzymes (*e.g.*, trypsin) to break the proteins down into shorter peptides. The peptides are then separated using High Performance Liquid Chromatography (HPLC) and injected into the mass spectrometer, where the peptides are fragmented into peptide fragments, ionised, and finally captured by the mass spectrometer. One experiment may generate thousands of MS/MS spectra, each of which theoretically corresponds to one of the proteins in the sample. However, mass spectra are usually tempered with noise and various artefacts. Thus the identification of proteins from mass spectra is a very challenging and error-prone process. Recent advances in mass spectrometry instruments and new fragmentation technologies provide unprecedented resolving power and mass accuracy in acquired spectra, which present a new opportunity to potentially identify 100% of the proteins and many more protein modifications than before [2 - 4]. However, with existing identification methods, only 50% of the proteins can be successfully identified and the protein post-translational modifications (PTM) are virtually unidentifiable [5 - 8]. Therefore, it has become a serious bottleneck for proteomics research and there is a critical need for more accurate protein identification methods that can fully utilise the resolving power of new instruments and identify more proteins and protein modifications.

Existing identification methods may be roughly classified into two categories: the database search approach and the de novo sequencing approach. The database search approach has been widely used due to its accuracy and reliability. Database search methods identify proteins by generating theoretical spectra *in silico* from a given protein database and comparing the experimental spectra with the theoretical spectra to find the best match. The main difference between database search methods lies in the type of scoring functions utilised to rank-order the most probable protein matches. One popular scoring method is exemplified by the SEQUEST algorithm [9], which applies a signal processing technique known as cross correlation to mathematically determine the overlap between the theoretical spectra and the experimental spectra to find the best match. Another important scoring method is to employ a probability model to estimate the likelihood of a match between the experimental spectrum and the theoretical spectrum being a random event. A number of methods have been proposed using such an approach, including X!Tandem [10] which uses a hyper-geometric model, OMSSA [11] which applies a Poisson model, and MASCOT [12]. It is very desirable that the probability-based database search methods provide direct measurement of the statistical confidence of an identified protein.

Despite the sophistication of database search methods, they have several limitations. Firstly, they are only effective if the proteins of interest are already known and the database used in the identification process contains the correct protein sequences. Unfortunately, for many scenarios this is difficult since many studies involve unknown proteins or proteins that have not been completely annotated [13]. Secondly, the database search methods have limited capability in detecting protein modifications. If the proteins in the samples are heavily modified, it usually leads to incorrect identifications for database search methods [14, 15]. Thirdly, specifying the enzyme used in the proteolytic digestion can also exclude the correct peptides from the search space and lead to misidentifications [16]. The de novo sequencing approach on the other hand is able to address these issues because it identifies proteins by extracting protein sequence information directly from experimental spectra and does not require any protein database. De novo sequencing methods are the only feasible means for applications such as finding novel proteins, detecting amino acid mutations, studying the proteome at the same time as the genome, and so on. However, the main obstacle for the de novo sequencing approach is that it usually requires relatively higher quality spectra. The recent development of mass spectrometry instruments enables the measurement of high dimensional mass spectra and provides unprecedented mass accuracy, and this has removed the main obstacle for the de novo sequencing approach.

Two different de novo sequencing methods have been developed. The first method, such as Sherenga [17] and Lutefisk [18], projects the problem into graph theory and applies algorithms for finding maximum path lengths in a network topology to achieve protein identification. The second method applies probability models in inferring the proteins from the spectra, for example NovoHMM [19] and PepNovo [20]. However, the main idea of these two methods is the same: to find the longest possible peptide sequence that best suits the observed experimental spectrum. Because many peaks in the spectra corresponding to real peptide fragment ions cannot be detected in the presence of protein modifications, and ion degradation generates many intensive peaks that cannot be explained, the optimal path may not always be the correct peptide identification. Therefore, we propose a new Bayesian network and dynamic programming hybrid de novo sequencing method to infer the most likely peptide sequences by exploring the sub-optimal space. The method firstly applies a Bayesian network probability model to infer a number of most probable peptide sequences given the spectra, and then utilises a dynamic programming algorithm to find the most likely sequence. Evaluated on a large number of tandem mass spectra, our method is able to outperform the most popular de novo sequencing algorithms.

II. METHOD

A. Terminology

A peptide P which has n amino acids can be formalised as: $P = p_1p_2\ldots p_n$. The total mass of the peptide therefore can be formalised as:

$$M = \sum_{i=1}^{n} m_i + 18, \tag{1}$$

where m_i is the residue amino acid mass, and 18 is the mass of H_2O. When peptides are subjected to fragmentation, a typical event is a single cleavage along the peptide's backbone. For an n amino acids peptide, there will be n possible cleavage positions, including the case that no cleavage happens. As a result, a peptide may result in a series of different ions based on the cleavage position. The N-terminal fragments (also called prefix fragments) can be denoted as: p_1, p_2… p_i, and the C-terminal fragments (suffix fragments) are then denoted as: p_{i+1}, … p_n. These peptide fragments will generate corresponding fragment ions with positive charges after ionisation, and a tandem mass spectrum is the collection of all detected signals of generated peptide fragment ions. N-terminal ions are called a-, b-, and c-ions, while the C-terminal ions are called x-, y-, and z-ions. If a cleavage happens at the i^{th} peptide bond, it will produce a_i, b_i, c_i ions and x_{n-i}, y_{n-i}, z_{n-i} ions. An illustration of possible peptide fragmentation positions and corresponding notations for the fragment ions is given in Figure 1. The peptide fragment ions may also have neutral losses, where chemical groups such as water or ammonia (NH_3) are separated from the fragment ions.

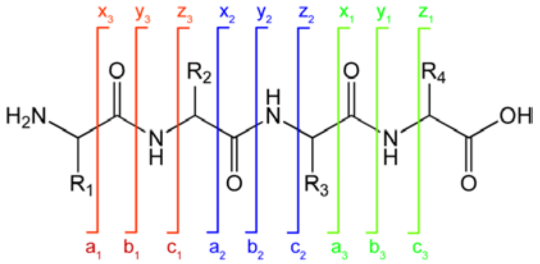


Figure 1. An illustration of a 4 amino acids peptide fragmentation pattern and notation for the fragment ions.

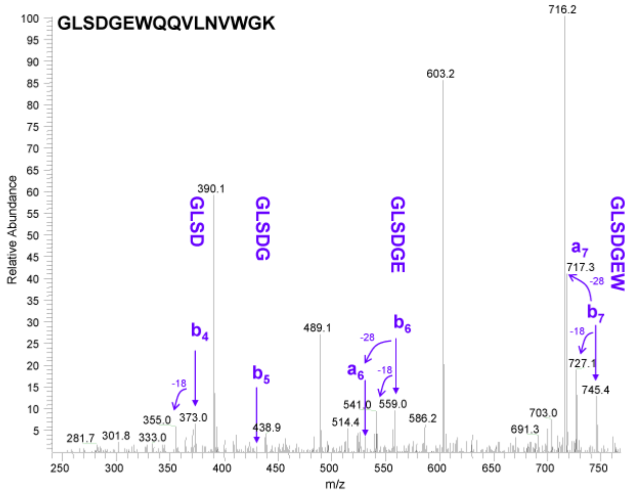


Figure 2. An example of identifying a peptide from a tandem mass spectrum using a de novo sequencing approach. The peptide precursor is singly charged and the spectrum is generated from an ion-trap mass spectrometer.

The mass spectrum of one peptide is a list of pairs of mass to charge ratio (m/z) and an associated intensity (m_1, i_1), (m_2, i_2), … (m_j, i_j) known as peaks, coupled with a parent (also called precursor) peptide mass M. The de novo sequencing problem is to infer the sequence of the peptide that gives rise to these peaks. Ideally each peak corresponds to one fragment ion, and the peptide sequence can be inferred from the mass difference between two adjacent peaks. An example is given in Figure 2. This is a very difficult task in reality, because spectra are very noisy and of complex nature. In addition, different fragment ions are not detected at the same probability and many fragment ions are hardly distinguishable from the background spectrum noise. For example, the signals of b- and y-ions may be up to 5 times stronger than those of a- and x-ions; 1/5 of the b- and y-ions may suffer from neural losses; z-ions usually have very low intensities and so on [9].

B. *Step 1: Spectrum Preprocessing*

Our method has three major steps: (1) spectrum preprocessing, (2) Bayesian network-based identification, and (3) inferring the most likely sequence. The first step is to preprocess the spectra peaks and normalise the peak intensities prior to the main de novo sequencing algorithm. Our method adopts the peak preprocessing procedure of PepNovo [20]. The method firstly determines the baseline intensity as the average intensity of the weakest 1/3 of the peaks in the spectrum. The method divides each peak's intensity to the baseline intensity so that a normalised intensity is obtained. The normalised peak intensities are discretised into 4 levels: no signal, low signal, medium signal, and strong signal. The method then removes the low signal peaks by sliding a window of width h across the spectrum and removing all the peaks except the top k peaks. For our method, we use $h = 15$ Da and $k = 3$. The method also constrains the total number of selected peaks to be no more than 100.

Because different regions of the spectrum have different characteristics and distributions of the peaks, our method organises the peaks into 5 regions based on their m/z positions and adds this information to the Bayesian network-based model. Therefore, the correlation between the peptide fragmentation and the observed peak intensities can be better captured. For example, peaks are usually more intensive in the middle region of the spectrum because peptides are less likely to be cleaved at the positions near the two termini.

C. *Step 2: Bayesian Network Identification*

The second step is to infer a number of most probable peptide sequences using a Bayesian network probability model. This step involves 4 procedures.

Procedure 1: The method constructs a spectrum graph as introduced in [17]. A spectrum graph is a directed acyclic graph, whose vertices correspond to putative ions of the peptide fragmentation. Two vertices are connected by a directed edge from the vertex with a lower mass to the one with a higher mass if the mass difference between these two vertices approximates the residue mass of an amino acid or other mass offsets like ion neural losses (see Table 1 for the complete list of all considered mass offsets). Given a preprocessed mass spectrum S, we build the entire spectrum graph and connect all the edges given the peaks of S. Since the most intensive peaks in the spectrum tend to be b- and y-ions, our spectrum graph has vertices for both interpretations: given a peak at mass m_i, we create a vertex at mass $m_i - 1$ interpreting the peak as a b-ion and also a vertex at mass $M - m_i + 1$ interpreting the peak as a y-ion, where M is the sum of residue amino acid masses. A vertex for an empty peptide of mass zero and a vertex for intact peptide of mass $M - 18$ are also added to the graph. If vertices are too close to each other (mass difference < 0.5 Da), these peaks are likely to be isotopic peaks of the same ion and are therefore merged. DiMaggio and Floudas [16] gave visualisation of a spectrum graph (also see Figure 3).

TABLE I. THE LIST OF ALL THE FRAGMENTIONS THAT ARE MODELLED; M IS THE SUM OF THE AMINO ACID RESIDUE MASSES.

Ion Type	**Notation**	
	Mass offset	**Terminus**
b^+	M + 1	C-Terminus
b^+ - H_2O	M – 17	C-Terminus
b^+ - NH_3	M – 16	C-Terminus
b^+ - $2H_2O$	M – 35	C-Terminus
b^+ - NH_3 - H_2O	M – 34	C-Terminus
b^{2+}	(M + 2)/2	C-Terminus
a^+	M – 27	C-Terminus
a^+ - H_2O	M – 45	C-Terminus
a^+ - NH_3	M – 44	C-Terminus
y^+	M + 19	N-Terminus
y^+ - H_2O	M + 1	N-Terminus
y^+ - NH_3	M + 2	N-Terminus
y^+ - $2H_2O$	M – 17	N-Terminus
y^+ - NH_3 - H_2O	M – 16	N-Terminus
y^{2+}	(M + 20)/2	N-Terminus

Procedure 2: Our method uses a Bayesian network model to calculate the probability of observing each vertex of the constructed spectrum graph. We adopted the fragmentation model proposed in [20] which incorporates several ion degradations (given in Table 1) and 3 additional factors into the model. These 3 factors are: (1) the relationship among different types of fragment ions; (2) the correlation between peptide cleavage position and the fragmentation efficiency; and (3) the influence of the last amino acid that is adjacent to the peptide terminus. Factor 1 models the strong correlation among a-, b- and y-ions. For instance, if a b-ion is detected, it is very common that its corresponding y-ion can be detected with high intensities, and its associated a-ion is usually detected. Although all ions have correlations, only a-, b- and y-ions regularly have strong signals therefore our method focuses on these ions. Factor 2 models that ions have different probabilities of being observed depending on the cleavage positions. For example, a-ions tend to be observed more often near the N-terminus, while b- and y-ions show much higher intensities in the middle region of the spectrum, and so on. Factor 3 models the N-terminal and C-terminal amino acids' chemical effects on the peptide cleavage as reported in the literature [21, 22]. The rest of the vertices model the probabilities of observing ion degradations and ions carrying multiple charges. The whole Bayesian network

is given in Figure 4. Except for the top 3 vertices, which represent the 3 additional factors, each vertex of the network contains a conditional probability table given the values of its parent vertices. For instance, if we use the second red path in Figure 4, vertex y+ holds the probability table $P(y^+ = t_i \mid b^+ = t_j$, region($i$) = R_k, NT(i - 1 or i + 1) = {any AA}, CT(i - 1 or i + 1) = {any AA}), where t_i is the intensity of the y^+ ion, t_j is the corresponding intensity of b^+ ion, R_k is the cleavage region of the spectrum, and NT and CT are the effects of the adjacent N-terminal and C-terminal amino acids respectively.

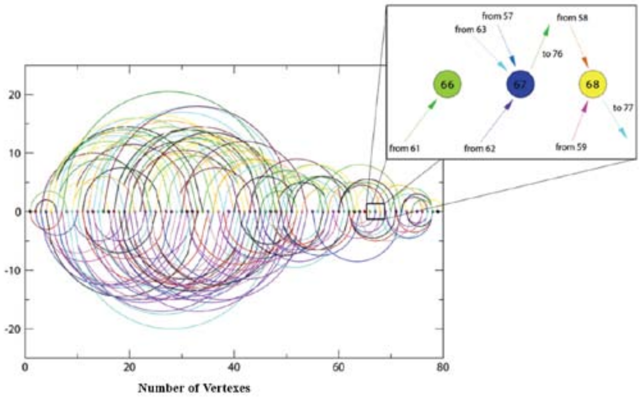


Figure 3. Visualisation of one spectrum graph, where each vertex represents one possible peptide cleavage position, and one directed edge is added if the mass difference between 2 vertices approximates the mass of an amino acid or an ion neutral loss.

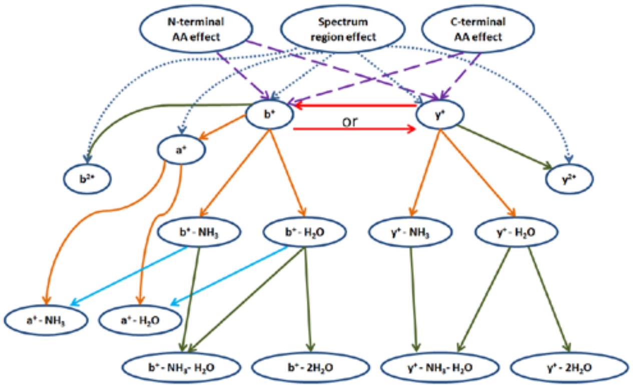


Figure 4. The Bayesian network model for our method. One of the two red paths will be randomly selected. The probabilities of the fragment ions are indicated by the colour of the solid paths: red > yellow > light blue > green. The dash paths are the additional 3 factors.

Our model differentiates itself from the one proposed in [20] in that it extends the way the probability tables are generated by incorporating singly charged, doubly charged, and triply charged tandem mass spectra from a mixture of mass spectrometry instruments. The Seattle dataset [23], which contains spectra from both ion-trap and quadrupole time of flight (TOF) mass spectrometers, was used for estimating the probability tables. More details are given in Section III. In this way, the model becomes more robust and can be applied to a much wider range of experiments.

Procedure 3: Each vertex of the constructed spectrum graph is scored using the described Bayesian network. This is achieved by comparing one hypothesis that the peak is a real fragment ion to the other hypothesis that the match is random. It is calculated by the likelihood ratio given in Equation (2):

$$O_i(m_j, S) = \log \frac{P_{real}(t \mid m_j, S, R, NT, CT)}{P_{random}(t \mid m_j, S)}, \tag{2}$$

where O_i represents the score for vertex i, m_j is the mass of the peak, S is the mass spectrum, t is the complete set of all peak intensities of S, R is the peak region, NT represents the N-terminal amino acid's chemical effect, and CT represents the C-terminal amino acid's chemical effect. Assume V is the set of the vertices in the probability network except the top 3 vertices, then $V = \{b^+, y^+, b^+ - H_2O, y^{2+} \ldots\}$. For each vertex v of V, $w(v)$ denotes v's parents' assigned intensities given the network topology. $P_{real}(t_v = i \mid w(v) = \{t_1, t_2, \ldots\})$ is the probability of detecting intensity i at fragment ion v given the intensities detected at its parents. Because all the conditional probability tables of the network have been obtained through training the Seattle dataset and vertex v is to be independent of the other vertices given that the values of its parents are known, the probability of observing ion fragment intensities t given that the possible cleavage occurred at mass m_j in spectrum S can be calculated by Equation (3):

$$P_{real}(t \mid m_j, S) = \prod_{v \in V} P_{real}(t_v \mid w(v), m_j, S, R, NT, CT). \tag{3}$$

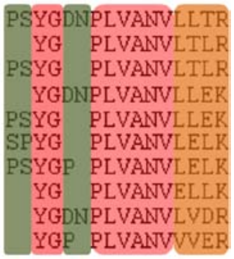
One advantage of the model is that P_{real} can distinguish the likely combinations of ions and ion degradations from unlikely combinations, since the conditional probability tables are learnt from real data. For example, the probability of observing a y^+ ion and its neural loss y^+ - NH_3 is higher than the probability of observing a y^+ - NH_3 ion without detecting the y^+ ion itself.

Under the hypothesis that the mass matches are random events, each peak is therefore considered to be independent. The probability of $P_{random}(t \mid m_j, S)$ can be easily calculated as the product of the probability of observing individual peaks at their mass positions. Once we have both P_{real} and P_{random}, the score for each vertex can be calculated.

Procedure 4: Given the spectrum graph and the score for each vertex, the method then finds several highest scoring asymmetric paths as the most probable peptide sequences. It is important to preserve the asymmetry because each peak from the spectrum contributes to two vertices in the constructed spectrum graph since we model both b- and y-ions for each peak. Dynamic programming is able to solve this problem and finds the highest scoring maximum path that goes through every pair of vertices corresponding to the same peak at most once. However, it has been shown that the maximum path may not be the best solution [24, 25]. There are two reasons: (1) a certain number of vertices on the optimal paths may be false positives because many high intensive peaks in the spectrum are signals from various interferences, including protein modifications, unexpected peptide internal fragments, contaminations, *etc*; and (2) several vertices representing the real peptide fragment ions may not have the highest score so will not be included in the optimal path. It is common that real fragment ions have low intensive signals or even cannot be detected at all. Therefore, we utilise the algorithm proposed by Lu and Chen [25] to obtain a set of most probable peptide sequences by exploring the sub-optimal solutions from the spectrum graph. The algorithm firstly transforms the spectrum graph into a matrix and uses an iterative depth-first search algorithm to find the optimal path. Sub-optimal solutions are obtained by back-tracking: at a certain iteration if a path showing close enough score to the optimal path, a sub-optimal path is then spawned and continued. Details of this algorithm can be found in [25].

D. Step 3: Inferring The Most Likely Sequence

The third step is to infer the most likely peptide sequence given the optimal sequence and a set of sub-optimal sequences. This set of peptide sequences has two main characteristics: (1) the majority of these sequences will have identical or highly similar segments of sequences; and (2) certain regions or sites may have ambiguities and show conflicting sequences. An example is given in Figure 5. The highly similar segments of sequences correspond to the high intensity fragment ions that are very likely to be correctly identified, while the ambiguous segments are where the peaks do not match fragment ions well or the intensities of the ions are hardly distinguishable from baseline noise. In addition, these sub-optimal solutions may have different numbers of amino acids.



```
PSYGDNPLVANVLLTR
  YG  PLVANVLTLR
PSYG  PLVANVLTLR
  YGDNPLVANVLLEK
PSYG  PLVANVLLEK
SPYG  PLVANVLELK
PSYGP PLVANVLELK
  YG  PLVANVELLK
  YGDNPLVANVLVDR
  YGP PLVANVVVER
```

Figure 5. A set of sub-optimal peptide sequences generated in Step 3. The red regions are the highly likely regions; the yellow region is the borderline region; the green regions are ambiguous regions.

Given these characteristics, the most likely peptide sequence can be extracted by adapting a dynamic programming-based algorithm similar to ClustalW [26] which has been used in multiple sequence alignment. In our case, the introduced “gaps” between the sub-optimal peptide sequences correspond to the ambiguous sections of the tandem mass spectrum. Our algorithm employs a progressive design and has 4 procedures in total.

Procedure 1: The pairwise distances of the sub-optimal peptide sequences are calculated using the Smith-Waterman dynamic programming algorithm [27]. An n by n distance matrix is then constructed from the pairwise distances, where n is the number of sub-optimal peptide sequences.

Procedure 2: A relationship for the sub-optimal peptide sequences is obtained given the distance matrix. The relationship is represented as a binary tree topology, and is constructed by applying the Neighbour Joining algorithm [28]. This algorithm is guaranteed to find the relationship topology that has the minimum overall distance.

Procedure 3: The peptide sequences are progressively aligned following the branching order of the constructed binary tree representing the relationship. The alignment proceeds from the tips of the relationship tree toward the root. In this way, the closest peptide sequences are aligned first, while the order of the most distant peptide sequences to be aligned is delayed.

Procedure 4: The final peptide sequence is obtained by identifying the highly likely segments of peptide sequences. Our method considers the regions highly likely if 85% or more of the peptide sequences agree on them. The most frequently appearing sequences will be used for these segments. The segments that are agreed by more than 55% (and less than 85%) of the sequences will be classified as borderline segments. Each amino acid in borderline segments will be determined based on its frequency across all the sub-optimal sequences. For example, if the frequencies for Glycine, Serine and Valine are 68%, 23% and 9% respectively at one site, then the algorithm will select one of these amino acids using the same probabilities as their frequencies. On the other hand, the “gaps” are interpreted as ambiguous sequence segments, which are denoted as undetermined “X” in the final identified peptide sequence.

III. RESULTS

A. Evaluation Strategy

As mentioned, we used the Seattle dataset [23] to learn the Bayesian network conditional probability tables. The Seattle dataset is a collection of reference mass spectra of 18 commercial purified proteins generated by several mass spectrometers. We selected singly charged, doubly charged, and triply charged spectra to learn the conditional probability tables. We ignored all the quadruply charged spectra because they are less common and usually of poor quality. We also excluded all the spectra generated by the MALDI TOF mass spectrometers, because spectra acquired from these machines have low resolution.

We compare the performance of our method with the most popular PepNovo and NovoHMM de novo sequencing methods by the criterion of identification accuracy. The identification accuracy is defined as the ratio of the number of correct amino acids to the number of identified amino acids. We use one large publicly available dataset to evaluate these 3 methods. The dataset is a collection of MS and MS/MS spectra of a mixture of 9 commercial purified proteins, generated by the Thermo Electron LTQ quadrupole linear ion-trap mass spectrometer. There are 3 technical replicas for this dataset, and in total the dataset contains 58,081 tandem mass spectra.

B. Evaluation Results

Our evaluation results are presented in Figure 6. Trypsin digestion was specified for running all 3 methods. Two amino acid pairs (Q and K), (I and L) are considered identical, since they have identical monoisotopic masses. Identification of either of these amino acids is considered correct. For example, if the peptide sequence is QFIER, the identifications such as QFLER and KFIER are all considered to be correct. PepNovo and NovoHMM were executed at default parameters. For our method, we used error tolerance of 0.1 Da and the maximum number of sub-optimal solutions that are explored to generate the final result was set to 20, which seemed to produce the best results.

PepNovo and NovoHMM seem to have similar overall performance in terms of identification accuracy. However, NovoHMM tends to have slightly higher accuracy in identifying short length peptide sequences. As shown in Figure 6, NovoHMM outperforms PepNovo at sequence lengths from 3 to 6 amino acids; while PepNovo starts to display better accuracy than NovoHMM for sequence length of 7 and onward. This may be due to NovoHMM's Hidden Markov model beginning to overfit when the spectra are more complicated. In any case, the performance difference between these two methods is quite small.

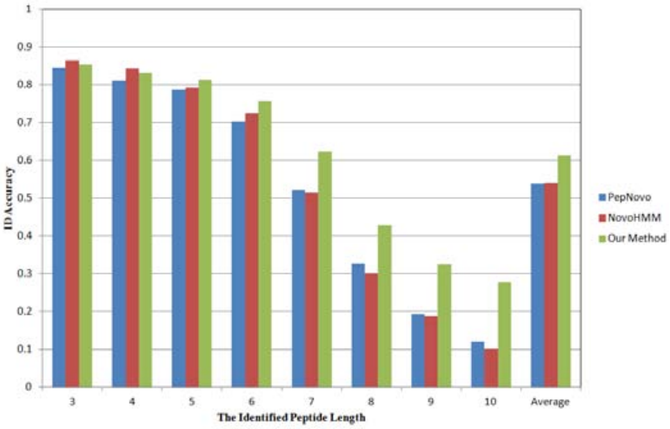


Figure 6. The comparison of identification accuracy. The x-axis is the identified peptide length in number of amino acids, the y-axis is the accuracy. The blue bar is PepNovo, the red is NovoHMM and the green is our method. The last 3 bars at the right end of the graph are the average accuracy across all peptide lengths.

Our method, compared to PepNovo and NovoHMM, has significantly better performance. It can be clearly seen from Figure 6 that our method on average achieved around 10% higher accuracy than PepNovo and NovoHMM. It is very promising that our method has much better accuracy in identifying peptide sequences of more than 7 amino acids. This is important because the majority of the tryptic peptides have 7-13 amino acids. Our evaluation results also indicate that our method has increasingly higher accuracy for longer peptides. Figure 6 shows that the accuracy improvement of our method at length 5 is minor, then it keeps increasing, and becomes almost doubled at peptide lengths of 9 and 10. This is probably because our method is not constrained to the maximum path and takes advantage of sub-optimal solutions. When peptides have more amino acids, the number of observed fragment ions may grow very quickly. Therefore, the likelihood that the optimal path is the correct peptide sequence becomes smaller and smaller. The results demonstrate that the exploration of sub-optimal space can significantly improve the identification accuracy.

IV. Discussion and future work

De novo sequencing based protein identification methods are commonly considered by the community as inferior to database search methods. This might be true for older MS instruments but is not the case anymore. Database search methods may be the first choice for low resolution spectra generated by older instruments; however database search methods render useless the resolving power of the new instruments. The identification coverage of database search methods simply cannot be significantly improved by using high resolution spectra. This is due to their reliance on protein databases, which are seldom complete. The de novo sequencing approach on the other hand is able to make better use of the high resolution spectra from new instruments and does not suffer from the issues of the database search approach. From our experiments, de novo sequencing methods are able to outperform typical database search methods on high resolution Orbitrap spectra data (results not shown). Therefore, the applicability of the de novo sequencing approach should be reconsidered and more research effort should be devoted to the development of new de novo sequencing methods.

Due to the complicated nature of mass spectra, not only the optimal solution but also the sub-optimal solutions should be utilised in order to improve the identification accuracy. Several de novo sequencing methods have been developed, all of which apply sophisticated algorithms. However, the central dogma of these methods remains the same: to find the maximum path in a spectrum graph under a specific model. Unfortunately, the optimal solution may not always be the correct identification. There are several explanations. Firstly, a large portion of highly intensive peaks in the spectra are not the expected signals from peptide fragment ions. This may be due to various reasons, such as peptide internal fragmentation, peptide post-translational modifications, contamination, chemical reactions, isotopic interferences, machine error, and many others. Secondly, many fragment ions are difficult to detect and usually have low intensities, for example c- and z-ions are barely distinguishable from noise. It is possible that even the dominant b- and y-ions are partially missing from the spectra. In any case, the fragmentation patterns still are not fully understood today. Therefore, the sub-optimal solutions are of great interest. The performance of our method clearly demonstrates that in a large number of cases the correct peptide sequences are not the optimal solutions, but can be obtained by exploring the top ranking sub-optimal solutions. This creates a new research direction and it would be very desirable to develop more efficient algorithms for exploring the sub-optimal space for accurate peptide identification.

The de novo sequencing approach has great potential for identifying protein modifications. One major advantage of our method is its ability to find the regions where the spectrum is difficult to explain. Many identified ambiguous regions turn out to be the locations where modifications tend to occur, especially phosphorylation. This is very interesting since phosphorylation is one of the most important protein modifications. It has been shown to activate or deactivate many protein enzymes and play key roles in cellular processes. This also indicates that protein modification is one important factor that greatly influences the accuracy of the de novo sequencing based identification. Although the identification of protein modifications is not the central concern of de novo sequencing, it remains the most effective approach because it infers the actual peptide sequences directly from the spectra rather than matching a database. If the de novo sequencing method has an efficient protein modification model, multiple protein modifications can be identified accurately by exploring the sub-optimal space. Our method may be easily extended for this purpose by incorporating further consideration of protein modifications into the Bayesian network, and this would be an interesting direction for future research.

Acknowledgment

This research was funded by Australian National Health and Medical Research Council (NHMRC) grant 525453. We adapted source codes from PepNovo for constructing the spectrum graph and probability model. We adapted source codes from ClustalW for generating the final peptide sequence.

References

[1] Zhang Q., Faca V., and Hanash S., “Mining the plasma proteome for disease applications across seven logs of protein abundance”, *J. Proteome Res*., vol. 10, pp. 46-50, 2011.

[2] Rinner O., et al. “Identification of cross-linked peptides from large sequence databases”, *Nat. Methods*, vol. 5, pp. 315-318, 2008.

[3] Tran J.C., et al. “Mapping intact protein isoforms in discovery mode using top-down proteomics”, *Nature*, vol. 480, pp. 254-258, 2011.

[4] Durbin K.R., et al. “Intact mass detectionm interpretation, and visualization to automate top-down proteomics on a large scale”, *Proteomics*, vol. 10, pp. 3589-3597, 2010.

[5] Spirin V., et al. “Assigning spectrum-specific P-values to protein identifications by mass spectrometry”, *Bioinformatics*, vol. 27, pp. 112801134, 2011.

[6] Deutsh E.W., Lam H., and Aebersold R., “Data analysis and bioinformatics tools for tandem mass spectrometry in proteomics”, *Physiol. Genomics*, vol. 14, pp. 18-25, 2008.

[7] Polaco, B.J., et al. “Discovering mercury protein modifications in whole proteomes using natural isotope distributions observed in liquid chromatography-tandem mass spectrometry”, *Mol. Cell Proteomics*, vol. 10, Epub 2011 Apr 30, 2011.

[8] Nesvizhskii A.I. and Aebersold R., “Analysis and validation of proteomic data generated by tandem mass spectrometry”, *Nat. Method*, vol. 4, pp. 787-797, 2007.

[9] Eng J.K, McCormack, A.L., and Yates J.R. “An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database”, *J. Am. Soc. Mass Spectrom.*, vol. 5, pp. 976-989, 1994.

[10] Craig R., Beavis R.C. “TANDEM: matching proteins with tandem mass spectra”, *Bioinformatics*, vol. 20, pp. 1466-1467, 2004.

[11] Geer L.Y., et al. “Open mass spectrometry search algorithm”, *J. Proteome Res*., vol. 3, pp. 958-964, 2004.

[12] Perkins D.N., Pappin D.J.C., Creasy D.M., and Cottrell J.S. “Probability-based protein identification by searching sequence databases using mass spectrometry data”, *Electrophoresis*, vol. 20, pp. 3551-3567, 1999.

[13] Lu B., and Chen T. “Algorithms for de novo sequencing using tandem mass spectrometry”, *Biosilico*, vol 2, pp. 2, 2004.

[14] Searle B.C., et al. “Identification of protein modifications using MS/MS de novo sequencing and the OpenSea alignment algorithm”, *J. Proteome Res.*, vol. 4, pp. 546-554, 2005.

[15] Bandeira N., Tsur D., Frank A., and Pevzner P.A. “Protein identification by spectral networks analysis”, *Proc. Natl. Acad. Sci*., vol. 10, pp. 6140-6145, 2007.

[16] DiMaggio P.A., and Floudas, C.A. “De novo peptide identification via tandem mass spectrometry and integer linear optimisation”, *Anal. Chem.* vol. 79, pp. 1433-1446, 2007.

[17] Dancik, V., et al. “De novo peptide sequencing via tandem mass spectrometry*”, J. Comput. Biol*., vol. 6, pp. 327-342, 1999.

[18] Taylor J. A., and Johnson, R.S. “Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry”, *Anal. Chem.*, vol. 73, pp. 2594-2604, 2001.

[19] Fischer B., et al. “NovoHMM: a hidden Markov model for de novo peptide sequencing”, *Anal. Chem*., vol. 77, pp. 7265-7273, 2005.

[20] Frank A. and Pevzner P. “PepNovo: de novo peptide sequencing via probabilistic network modelling”, *Anal. Chem*., vol. 77, pp. 964:973, 2005.

[21] Tabb D.L., Smith L.L., Breci L.A., Wysocki V.H., Lin D., and Yates J.R. “GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model”, *Anal. Chem*., vol. 75, pp. 1155-1163, 2003.

[22] Breci L.A., Tabb D.L., Yates J.R., and Wysocki V.H., “Cleavage N-terminal to Proline: analysis of a database of peptide tandem mass spectra”, *Anal. Chem*., vol. 75, pp. 1963-1971, 2003.

[23] Limek J., et al. “The standard protein mix database : a diverse dataset to assist in the production of improved peptide and protein identification software tools”, J. Proteome Res., vol. 7, pp. 96-103, 2008.

[24] Chen T., Kao M.Y. “A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry”, J. Comput. Biol., vol. 8, pp. 325-337, 2001.

[25] Lu B. and Chen T.J. “A suboptimal algorithm for de novo peptide sequencing via tandem mass spectrometry”, J. Comput. Biol., vol. 10, pp. 1-12, 2003.

[26] Thompson J.D., Higgins D.G., and Gibson T.J., “CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specofic gap penalties and weight matrix choice”, Nucleic Acids Res., vol. 22, pp. 4673-4680, 1994.

[27] Smith T.F. and Waterman M.S., “Identification of common molecular subsequences”, *J. Mol. Biol*., vol. 25, pp. 195-197, 1981.

[28] Saitou N. and Nei M., “The neighbor-joining method: a new method for reconstructing phylogenetic trees”, *Mol. Biol. Evol*., vol. 4, pp. 406-425, 1987.

[29] Thompson J.D., Higgins D.G., and Gibson T.J., “CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice”, *Nucleic Acids Res.*, vol. 22, pp. 4673-4680, 1994.