Sequence analysis

**De novo** peptide sequencing using ion peak intensity and amino acid cleavage intensity ratio

Mitsuhiro Kanazawa¹,5,* , Hisae Anyoji¹, Atsushi Ogiwara²,³ and Umpei Nagashima⁴,5

¹Reifycs Inc., ²Medical ProteoScope Co., Ltd, ³Tokyo Medical University, Tokyo, Japan, ⁴National Institute of Advanced Industrial Science and Technology and ⁵Tsukuba University, Ibaraki, Japan

Received on September 17, 2006; revised on February 1, 2007; accepted on February 18, 2007

Advance Access publication March 6, 2007

Associate Editor: Martin Bishop

**ABSTRACT**

**Motivation:** Peptide-sequencing methods by mass spectrum use the following two approaches: database searching and **de novo** sequencing. The database-searching approach is convenient; however, in cases wherein the corresponding sequences are not included in the databases, the exact identification is difficult. On the other hand, in the case of **de novo** sequencing, no preliminary information is necessary; however, continuous amino acid sequence peaks and the differentiation of these peaks are required. It is, however, very difficult to obtain and differentiate the peaks of all amino acids by using an actual spectrum.

We propose a novel **de novo** sequencing approach using not only mass-to-charge ratio but also ion peak intensity and amino acid cleavage intensity ratio (CIR).

**Results:** Our method compensates for any undetectable amino acid peak intervals by estimating the amino acid set and the probability of peak expression based on amino acid CIR. It provides more accurate identification of sequences than the existing methods, by which it is usually difficult to sequence.

**Contact:** kanazawa@reifycs.com

1 INTRODUCTION

DNA sequences of various organisms have been decoded in recent years. However, the sequence itself does not contribute directly to biological functions until it is translated into a protein. Therefore, the study of all the proteins expressed in a cell, which is commonly known as ‘proteomics’, is currently being used as an approach to elucidate disease mechanisms and clinical conditions. The DNA sequencer is a well-known instrument used for decoding DNA sequences; moreover, the outstanding achievement of the human genome project can be attributed to the use of this instrument for decoding DNA sequence fragments. On the other hand, practical protein-sequencing instruments or methods, in short, ‘peptide sequencing’, are yet to be established as DNA-sequencing methods, and the establishment of peptide-sequencing methods is essential for future medical research and development.

Currently, mass spectrometric techniques are being increasingly used for peptide-sequencing methods. Remarkable progress has been achieved in peptide ionization and fragmentation techniques by K. Tanaka and J.B. Fenn, for which they received the 2002 Nobel Prize in Chemistry. Their techniques employ soft ionization that does not damage peptides but, instead, produces various ion fragments with few peptide dissociation sites. In mass spectrometry, two types of data are measured: one spectrum indicates the weights of the peptides (MS), and the other spectrum provides the weight of the peptide fragment created by the dissociation of each peptide (MS). The exponent ‘n’ represents (n=1)th peptide dissociation. In this article, both MS² and MS³-piled spectra derived from a peptide have been reported. (Unless specified otherwise, hereafter, MS² implies both MS² and MS³-piled spectra derived from a peptide.)

Peptide-sequencing methods involving the MS² spectrum employ the following two approaches: database searching and **de novo** sequencing. Database searching methods, for example, MASCOT (Perkins et al., 1999), are convenient; however, in cases wherein the corresponding sequences are not included in the databases, the identification of the exact amino acid sequence is difficult. In the case of **de novo** sequencing, for example, PEAKS (Ma et al., 2003), no preliminary information is necessary; however, continuous amino acid sequence peaks and the differentiation of these peaks from the other peaks are required. It is, however, very difficult to obtain the peaks of all amino acids along with their differentiation from the other peaks by using an actual spectrum.

In this study, we propose a novel **de novo** sequencing method to extract sequences without the use of databases by assuming the existence of peaks of all amino acids. The existing **de novo** peptide-sequencing approaches use only the peak interval of mass-to-charge ratio (m/z) in the MS² spectrum; this spectrum consists of 2D values of m/z and ion intensity. On the other hand, our method uses not only m/z but also ion peak intensity and amino acid cleavage intensity ratio (CIR) (Kapp et al., 2003). Our novel method compensates for any undetectable amino acid peak intervals by estimating the amino acid set and the probability of peak expression based on amino acid CIR. Our method realized more accurate sequencing than the existing **de novo** sequencing methods, by which it is usually difficult to differentiate these sequences.

*To whom correspondence should be addressed.
2 MATERIALS

High ion intensity peaks indicate that a large number of dissociated fragments are detected as ions by mass spectrometry based on the dissociation energy produced within the instrument. This dissociation is not randomly produced. Three dissociation sites have been reported to exist, namely, dissociation at the peptide bond, dissociation at bonds preceding the peptide bond and dissociation at bonds following the peptide bond (Johnson et al., 1987). In other words, six fragments are generated by these three dissociations. Therefore, the values of fragment mass and ion intensity at these three dissociation sites are essential in order to identify peptide sequences. It is considered that the height of the ion intensity that indicates the amount of a fragment ion depends on the ease of dissociation of a combination between two amino acids and the abovementioned three dissociation sites.

With regard to the ease of dissociation between the amino acids, quantum chemical calculations or experimentally measured values of bond strings between the amino acids at the three dissociation sites might be more appropriate. However, previous studies have shown that the acquisition of these values is time consuming as it now stands. Therefore, in this study, we applied the statistical value of dissociation frequency as the ease of dissociation between the amino acids instead of the practical bond strength. These statistical values, for example, CIR, are determined based on the frequency of peak appearance at the dissociated site, regardless of the charge state that was obtained using the database method. (Equation (1)).

\[
\text{CIR}_{s} = \frac{\sum_{j=1}^{M} (b_{j}^{+} + y_{j}^{+})}{(1/N) \sum_{i=1}^{N} \sum_{j=1}^{M} (b_{j}^{+} + y_{j}^{+})}
\]

where, \(s\) is the dissociation site of the amino acid combinations, and \(j\) is the charge state of the fragment ion. This equation describes only the dissociation that is associated with the \(b\)-ion and \(y\)-ion.

3 METHODS

The algorithms of the existing methods are based on the comparison between the theoretical mass of amino acids and their observed mass determined by measurements. On the other hand, our method employs both the theoretical and observed masses as well as ion intensity to realize more practical sequencing. The core algorithms of our method use the following two parameters: one is the cost function that employs ion intensity to identify amino acids sequentially from one terminal to the other terminal, while the other is the amino acid set calculation that restricts the number of identified amino acid candidates.

3.1 Cost function employing ion intensity

Almost all existing methods, including the database method, calculate the mass difference \(M_{\text{Diff}}\) between the observed mass \(M_{\text{Observed}}\) and theoretical mass \(M_{\text{Theoretical}}\) with regard to dissociation sites as one of basic parameters for each algorithm (Equation (2)).

\[
M_{\text{Diff}} = |M_{\text{Theoretical}} - M_{\text{Observed}}|
\]

\(M_{\text{Diff}}\) is evaluated based on the complete identified sequence, and the minimized total \(M_{\text{Diff}}\) determines appropriate sequencing in the general method. Our cost function also employs the minimization of total \(M_{\text{Diff}}\); however, it will impose a penalty derived from the ion intensity on the minimized total \(M_{\text{Diff}}\) as well.

The CIRs corresponding to site \(s\) between amino acids and ion intensity \(I\) at \(s\) are defined by (Equation (3)).

\[
f(I) = \frac{g(CIR_{s})}{\text{Penalty}} = 1 \quad (3)
\]

where, \(f(x)\) is defined by the typical distribution of ion intensity obtained using various instruments and conditions, and \(g(x)\) is defined to obtain a constant value. Each \(f(x)\) and \(g(x)\) does not change any order of \(x\) values, and this equation adjusts the relationship between ion intensity \(I\) and CIR under specified conditions.

Based on this definition, the evaluation of relevance between \(I\) and CIR in the measured spectrum for sequencing is given as Penalty (Equation (4)).

\[
\text{Penalty} = \left| \frac{f(I)}{g(CIR_{s})} - 1 \right| \quad (4)
\]

It is clear that the minimization of total \(M_{\text{Diff}}\) involves the assignment of amino acid by using noise peaks, which are considerably difficult to avoid in actual measurements. In contrast, several low height peaks, such as undetectable peaks, have to be assigned as identified amino acids. Therefore, we propose the employing of the evaluation of ion intensity with the minimization of total \(M_{\text{Diff}}\) in the cost function (Equation (5)).

\[
\text{Cost} = \frac{1}{D} \sum_{i=1}^{D} M_{\text{Diff},i} \left| \frac{f(I)}{g(CIR_{s})} - 1 \right| \quad (5)
\]

\[
\text{Cost} = \frac{1}{D} \sum_{i=1}^{D} M_{\text{Diff},i} \left| \frac{f(I)}{g(CIR_{s})} \right| \quad (6)
\]

where \(D\) is the depth (number) of identified amino acids. The cost of the entire peptide sequence identification was evaluated by Penalty of the relationship between ion intensity \(I\) and CIR with \(M_{\text{Diff}}\).

Almost all existing methods filter the low ion intensity peaks prior to sequence identification in order to reduce calculation costs. However, this may affect sequencing because low CIR might not appear as a peak. On the other hand, the cost function can be used even when the ion intensity is low, for example, a high ion intensity peak is assumed to have a high CIR, and a low ion intensity peak is assumed to have a low CIR.

3.2 Amino acid set calculation

As mentioned earlier, the peak of certain fragments does not appear to depend on the combination of amino acids. The characteristic of our cost function is that it not only employs ion intensity but also considers undetectable peaks. However, unless the amino acids that apply \(M_{\text{Theoretical}}\) to cost function are defined, this function may not function appropriately for sequence identification.

The observed precursor mass or peak interval of MS* provides hints on the existing amino acids in the spectrum, even when certain peaks of amino acids are absent. Only a few conceivable amino acid sets exist, and the limitation of mass allows us to restrict the number of sets. We solved a knapsack problem by applying the Barnes–Hut tree code (Barnes and Hut, 1986) to list all suitable sets of amino acids against the masses. In the sequencing process using the cost function, if a difficulty is encountered in detecting any appropriate amino acid peaks, the possible sequence combination is calculated as an amino acid set and the evaluation is continued using this amino acid set. This continuation of sequencing using the amino acid set promptly
and accurately identifies the appropriate amino acid sequentially until the end of the sequence.

4 RESULTS
To illustrate our method, we used a peptide sequence ‘AEFVEVTK’ of bovine serum albumin (BSA); this peptide sequence was analyzed by Finnigan LTQ linear ITMS (Thermo Electron, San Jose, CA, USA) equipped with NSI sources (AMR Inc., Tokyo, Japan).

Figure 1 shows the identification process of this sequence by using search depth ($d$). This process involves the comparison of the theoretical mass of the fragment ion with its observed mass, calculation of the cost using the cost function and selection of appropriate amino acids as identified amino acids. Although the masses of leucine (L) and isoleucine (I) are the same, these amino acids have different costs due to different CIRs. These cost differences provide a possibility to differentiate sequencing results. The existing de novo peptide-sequencing methods are limited with respect to this differentiation. However, our method has the potential to identify the sequence itself along with the costs.

Let us assume that this identification process does not detect any amino acid peaks with significant cost after the identification of sequence under search depth $d = 3$. In the measurement experiments, the peaks of all amino acids in a peptide sequence may not always appear, and this assumption represents the typical difficulty encountered in peptide-sequence identification. In this case, the identification process calculates all the possible amino acid sets based on the remaining undetected sequence mass by using the knapsack problem-solving method. The calculated precursor mass is 923.036 Da; hence, the remaining undetected sequence mass is calculated to be 575.607 Da under $d = 4$. The amino acid set calculation process lists the possible sets based on 575.607 Da along with the weight margin. The identification process skips the undetectable peak of a certain amino acid once, and the cost function is simultaneously evaluated for this skipped amino acid belonging to the amino acid set and the next identified amino acid. The amino acid set calculation process continues sequence identification by temporal residue assignment under the limitation of the amino acid set until the other sequence terminal. Even when there are no peaks at $d = 4$, as mentioned earlier, this sequencing process calculates the cost to evaluate whether the amino acid at $d = 4$ is the correct amino acid.

![Fig. 1. The process of sequencing using the cost function.](image-url)
This is because cost is considered based on CIR, and low CIR implies less fragmentation.

After listing all the possible amino acid sequences, the identified amino acid sequences are ordered according to the cost: verisimilitude.

For the explanation of advantage using our method, we attempted amino acid sequence identification using both our present method and PEAKS Studio Demo Version 3.1. Processed spectra here are also peptide sequences of BSA, and we selected all 20 spectra that were identified independent sequences by Matrix Science Mascot Version 2.1 as a database approach. Table 1 shows sequence identification results of those 20 spectra by both methods and the hit order when each method identifies correct peptide sequence. Sequence listed in boldface indicates consensus sequence of 1st hit with correct identification.

Both methods provide us similar sequence identification candidates, but the order of candidates that are defined by scores of each algorithm is different. In Figure 2, the results of the amino acid sequence ‘AEFVEVT’ identification using the two methods are illustrated. PEAKS identified its sequence correctly, but it was fourth candidate. Our method appropriately identified the correct peptide sequences as 1st hit without considerable weight-difference permission of MDiff.

Table 1. Sequence identification results and hit order

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Present method</th>
<th>PEAKS</th>
</tr>
</thead>
<tbody>
<tr>
<td>AEFVEVT</td>
<td>AEFVEVT</td>
<td>1</td>
</tr>
<tr>
<td>AFDEK</td>
<td>AFDEK</td>
<td>1</td>
</tr>
<tr>
<td>DAFLGSFLYEYSR</td>
<td>ADFLGSFLYEYSR</td>
<td>2</td>
</tr>
<tr>
<td>DLGEEHFK</td>
<td>DLGEEHFK</td>
<td>1</td>
</tr>
<tr>
<td>DVLR</td>
<td>DVLR</td>
<td>1</td>
</tr>
<tr>
<td>HLVDPEQNL</td>
<td>No significant hit</td>
<td>–</td>
</tr>
<tr>
<td>HPYFPAELLYYANK</td>
<td>No significant hit</td>
<td>–</td>
</tr>
<tr>
<td>IETMR</td>
<td>IETMR</td>
<td>1</td>
</tr>
<tr>
<td>LFTHADICTLPOTEK</td>
<td>No significant hit</td>
<td>–</td>
</tr>
<tr>
<td>RGEYGFQNALIVR</td>
<td>LYGFGQNALIVR</td>
<td>1</td>
</tr>
<tr>
<td>LNELTEFAK</td>
<td>LNELTEFAK</td>
<td>1</td>
</tr>
<tr>
<td>LVTDLTK</td>
<td>LVTDLTK</td>
<td>1</td>
</tr>
<tr>
<td>QTALVELLK</td>
<td>QTALVELLK</td>
<td>1</td>
</tr>
<tr>
<td>SLHTLFGDELCK</td>
<td>No significant hit</td>
<td>–</td>
</tr>
<tr>
<td>TEEQLK</td>
<td>TEEQLK</td>
<td>1</td>
</tr>
<tr>
<td>TPVSEK</td>
<td>TPVSEK</td>
<td>1</td>
</tr>
<tr>
<td>TVNENPVAFVK</td>
<td>No significant hit</td>
<td>–</td>
</tr>
<tr>
<td>VASLR</td>
<td>VASLR</td>
<td>1</td>
</tr>
<tr>
<td>VPQVSTPTTLVEVSR</td>
<td>VPQVSTPTTLVEVSR</td>
<td>1</td>
</tr>
<tr>
<td>YLYEIQAR</td>
<td>No significant hit</td>
<td>–</td>
</tr>
</tbody>
</table>

5 CONCLUSION

Our peptide-sequencing method significantly improved the accuracy of the existing method by applying the following two features to peptide sequencing: (1) application of ion intensity using CIR and (2) application of amino acid set calculation.

The application of ion intensity to peptide sequencing facilitated the differentiation of sequences, which is usually difficult in the existing methods. In this study, we applied CIR as a statistical peptide cleavage frequency only on the b-ion and y-ion to employ ion intensity in the cost function. However, for using ion intensity, the study of peptide cleavage tendency as an alternative to CIR is important.

As PEAKS was able to perform database search after de novo sequencing, it identifies appropriate sequence as a final result in many case. On the other hand, the purpose of our present method is to detect novel sequences that are difficult to identify by database methods. Therefore, we plan to conduct more tests and compute accurate peptide cleavage tendency for further improvement using parameters such as statistical values, quantum chemical calculations or experimental measured values.
ACKNOWLEDGEMENTS

Implementation of this study was to discover novel peptide sequence under the clinical proteomic research at Medical Proteo Scope Co., Ltd, Tokyo, Japan. The authors gratefully acknowledge the technical advice of Dr Nobuhiro Fukushima of Science Technology Systems Inc., Tokyo, Japan, and the encouragement and assistance of Mr Kazunori Okamura, Hiraki & Associates, in the filing of patent application regarding this research.

We also thank Matrix Science Ltd and Bioinformatics Solutions Inc. for giving us opportunities to study peptide identification and sequencing algorithm.

Conflict of Interest: none declared.

REFERENCES


