Detection and Correction of Interference in MS1 Quantitation of Peptides Using their Isotope Distributions

Yifei Bao
Department of Computer Science
Stevens Institute of Technology

Beatrix Ueberheide
Department of Biochemistry and Molecular Pharmacology
and Proteomics Resource Lab
New York University Langone Medical Center

Manor Askenazi
The Ionomix Initiative, Arlington, MA

Joseph S. Glavy
Department of Chemistry, Chemical Biology and Biomedical Engineering
Stevens Institute of Technology

Jessica Chapman and David Fenyö
Center for Health Informatics and Bioinformatics
New York University Langone Medical Center

Department of Computer Science
Technical Report CS-2013-3
Abstract

Each peptide is observed as a group of peaks in mass spectrometry because of the presence of small amounts of stable heavy isotopes in nature. The natural isotopic distribution of a peptide is dependent on the number of atoms that make up the peptide and can be theoretically determined. Therefore, the isotopic distribution of a peptide is independent of peptide concentration, and any interference in MS1 quantitation of peptides can be detected by looking for deviations from a peptide’s theoretical isotopic distributions. In this study, computer simulations were used to build a library of different types and magnitude of interference, i.e. hypothetical interferences were added to theoretical isotope distribution for peptides. Interference correction was done through searching this library. The approach was demonstrated using experimental data from E. coli and S. cerevisiae tryptic digestion mixtures.

1 Introduction

The research and development of the Human Genome Project has provided a wealth of genetic information available to biologists. The major challenges now are to study the structure and function of the encoded proteins and to approach biological questions from the protein perspective to provide both qualitative and quantitative information about gene expression [1]. The structure and function of proteins are too complex to be predicted from analysis of nucleic acids alone. From transcription to translation, extensive changes can be introduced. Post-translational modifications (PTM) such as phosphorylation and dephosphorylation, glycosylation, and translocation within the cell can also dramatically change the protein products from a single gene. Proteomics is an emerging field to explore protein research from the overall level of cellular proteins [2–4].

A multitude of proteomic technologies have been developed to address the challenges from qualitative protein identification and quantitative analysis of changes in proteomes [5]. Two dimensional gel electrophoresis [6, 7] is the major tool to separate proteins on the basis of charge in the first dimension and molecular mass in the second. After separation, proteins can be cut into peptides by enzymes, which can be identified and quantified using mass spectrometry to measure their mass-charge ratio (m/z), and abundance in the sample. Mass spectrometry was originally developed almost 100 years ago to measure elemental atomic weights and the natural abundance of specific isotopes [8]. Over the pass decades, mass spectrometry (MS) has been the indispensable technology in proteomics for accurate identification and quantitation of proteins because of its sensitivity which allows for the quantitative measurement of low level peptides in complex mixtures [9].

In mass spectrometry based peptide quantitation, the measured signal is considered to have a linear dependence on the amount of material in the sample for the entire range of amounts being studied. Recent studies have been done to address the problems of signal processing and peak finding [10, 11]. The quantity of peptides is measured by calculating the height or the area of the corresponding peaks in the ion chromatograms. A challenging problem for peptide quantitation is that peptides with the same m/z will interfere with each other. The frequent occurrence of interferences is a significant problem that can result in inaccurate quantitation of peptides. However there are only a few recent studies that address interference detection and correction issues for MS2 quantitation, and there is very little focus on this problem for MS1 quantitation. Manual inspection is still typically used to identify interferences. To address the time-intensive and error-prone nature of manual inspection, we propose computational approaches here to detect and correct the interference in both MS1 quantitation of peptides and MRM quantitation of peptides.
2 Materials and methods

2.1 Theoretical isotope distribution calculation

The amount of stable heavy isotopes in nature is known, e.g., 0.015% $^2$H, 1.11% $^{13}$C and 0.366% $^{15}$N, 0.038% $^{17}$O, 0.200% $^{18}$O, 0.75% $^{33}$S, 4.21% $^{34}$S, and 0.02% $^{36}$S. The following equation (Equation 1) is usually used to calculate the intensity of peak $m$ in a mass spectrum caused by one specific heavy isotope.

$$T_m = \binom{n}{m} p^m (1 - p)^{(n-m)}$$

For instance, if the heavy isotope is $^{13}$C, $n$ is the number of carbon atoms in the peptide; $p$ is the probability of $^{13}$C in nature, which is 1.11%; $m$ is the number of $^{13}$C; and $T_m$ represents the probability of $m$ $^{13}$C atoms appear in the peptide. The intensity of the monoisotopic peak is considered to be 1, so that $T_m$ also represents the relative intensity of peak $m$ caused by the only heavy isotope $^{13}$C in the distribution.

The intensities of the theoretical isotope distribution of peptides are calculated by including all possible isotopes. We used Monte Carlo simulations to simulate the possible appearance of heavy isotopes in a peptide for each atom. The algorithm was repeated 10000000 times. The intensities is calculated by the normalizations between the count of peptides with heavy isotopes and the count of monoisotopic peptides. For each peptide, three peaks starting from the monoisotopic peak were used to represent the intensities of isotope distribution. The intensities of the three peaks of peptide GLPADVVPGDILLDGR are 1, 1.0147529010, and 0.5896354272.

2.2 Data collection

<table>
<thead>
<tr>
<th>Group</th>
<th>Mixture</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1µg E. coli tryptic digestion</td>
</tr>
<tr>
<td>2</td>
<td>1µg E. coli and 0.05µg S. cerevisiae tryptic digestions mixture</td>
</tr>
<tr>
<td>3</td>
<td>1µg E. coli and 0.1µg S. cerevisiae tryptic digestions mixture</td>
</tr>
</tbody>
</table>

Table 1. Samples for experimental data generation

In this study, three groups of samples as shown in Table 1 with different E. coli and S. cerevisiae tryptic digestion mixture ratios were prepared to generate experimental data. Those samples were analyzed using a Q Exactive (Thermo Scientific, Inc.) a hybrid quadrupole orbitrap mass spectrometer for peptide identification and quantitation. Group 1 has only E. coli tryptic digestion, which is a control group and supposed to have no interference from S. cerevisiae peptides. For the other two groups, some E. coli peptides would suffer interference from S. cerevisiae peptides. Our approach was applied to the experimental data for interference detection and correction. From the raw files, we extracted both the MS1 and MS2 datasets. We searched the data with X!Tandem (http://www.thegpm.org/tandem/index.html) using the MS2 dataset to identify a list of E. coli peptides and S. cerevisiae peptides. A Perl program was implemented to find the E. coli peptides which are interfered by the S. cerevisiae peptides though searching the peptides which may possibly have the same m/z value. GLPADVPGDILLDGR is a candidate E. coli peptide on the list, and it is used as an example here to interpret our approach.

The experimental data was uploaded to Slice (http://fenyolab.ionomix.com/), which is a web-based application for mass spectrometry data searching and visualization. The MS1 measured intensities of the peptides that we are going to analyze from the three different groups were obtained from Slice. Figure 1 shows the MS1 measurement around the highest intensity of monoisotopic peak of 2+ charged GLPADVPGDILLDGR. Each graph from the top to the bottom shows the data from pure E. coli.
sample (group 1), *E. coli* plus lower concentration of *S. cerevisiae* sample (group 2), and *E. coli* plus higher concentration of *S. cerevisiae* sample (group 3) respectively. On the right side of each panel, the intensities of theoretical isotope distribution are shown for this peptide.

### 2.3 Interference detection

Since the isotopic distribution of a peptide is independent of peptide concentration, the measured intensities in one specific peptides mass spectrometry without interference must be consistent with its theoretical isotope distribution. If there is interference, the intensities of one or more peaks will be increased and the changes can be detected by comparing the measured intensities with the theoretical isotope distribution. With obvious interference, manual inspection can be easily conducted by observing the plots as in Figure 1. However, manual inspection is error-prone and does not work for many cases. From Figure 1, observation with human eyes cannot help to draw the conclusion of whether there is interference for 2+ charged GLPADVVPGDILLDDGR in the samples, because the plots of measured intensities seem very similar to the theoretical isotope distribution. The measured intensities of 2+ charged GLPADVVPGDILLDDGR in the three difference samples can be found in Table 2 below. The 1st peak column represents the intensities of the monoisotopic peak. Compared with group 1, the measured intensities in group 2 and group 3 increase as the concentration of *S. cerevisiae* tryptic digestion increases, which indicates group 2 and group 3 suffer from interference based on the condition that group 1 has no interference from *S. cerevisiae* peptides.

We present here a computational approach to detect interference by checking whether measured isotopic distribution of a peptide deviates from the theoretical isotope distribution. We used the relative ratios between the intensities of different isotopic peaks as the signature features to define the isotope distribution. Equation 2 below calculates the difference between two isotope distribution. \(i\) is the number of peaks starting from monoisotopic peak, whose number is 1. \(I_{1,i}\) is the intensity of peak \(i\) from the measured isotope distribution and \(I_{2,i}\) is the intensity of peak \(i\) from the theoretical isotope distribution. \(\text{dif}_i\) represents the difference between two isotope distributions by comparing the relative ratios calculated by the first two peaks. We only focused on the first three peaks for each measurement. Hence, \(i\) can be 1 or 2. We set the threshold for the difference between two isotope distributions to be 0.1, which means if either \(\text{dif}_1\) or \(\text{dif}_2\) is greater than 0.1, we consider that the measurement suffers from interference.

\[
\text{dif}_i = \frac{\text{abs}(I_{1,i+1} - I_{2,i+1})}{I_{2,i+1}} (i = 1, 2, 3, \text{if } I_{1,i} \neq 0, I_{2,i} \neq 0, \text{and } I_{2,i+1} \neq 0)
\]

As mentioned in Section 2.1, the intensity ratios of the three peaks starting from monoisotopic peak for peptide GLPADVVPGDILLDDGR is theoretically to be 1: 1.0147529010: 0.5896354272. Table 2 below shows the calculation of \(\text{dif}_1\) and \(\text{dif}_2\) for the measurements of 2+ charged GLPADVVPGDILLDDGR. Group 2 and 3 are detected to have interference based on the calculation, and the result is consistent with the experimental data.

<table>
<thead>
<tr>
<th>Group</th>
<th>1st Peak</th>
<th>2nd Peak</th>
<th>3rd Peak</th>
<th>(\text{dif}_1)</th>
<th>(\text{dif}_2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4101069</td>
<td>3871992</td>
<td>2037097</td>
<td>0.06958512</td>
<td>0.090449</td>
</tr>
<tr>
<td>2</td>
<td>6136956</td>
<td>5574542</td>
<td>3019380</td>
<td>0.104850</td>
<td>0.0636069</td>
</tr>
<tr>
<td>3</td>
<td>7720741</td>
<td>7526431</td>
<td>3893232</td>
<td>0.0393407</td>
<td>0.105725</td>
</tr>
</tbody>
</table>

**Table 2.** Interference detection based on the experimental data for 2+ GLPADVVPGDILLDDGR. The threshold was set to 0.1, and both group 2 and group 3 were detected to have interference, while group 1 had no interference detected.
Figure 1. MS1 intensity distribution for 2+ charged GMPADVVGDILLDGR. Black represents the first peak which is the monoisotopic peak, red represents the second peak and yellow represents the third peak. The intensities of the theoretical isotope distribution are presented on the right side of each panel. X axis is the scan time, and y axis is the intensity.
2.4 Interference library construction and interference correction

Compared with interference detection, it is a much more challenging task to correct the interference, because it is necessary to know the intensities of which peaks are increased by interference and how much is increased. We propose an approach to simulate the different types and magnitude of interference that one peptide can have which constructs a library of mass spectrum patterns. For each pattern in the library, both the theoretical isotope distribution of the peptide and what kind of interferences was added to the theoretical isotope distribution are known information. The interference correction can be done by searching matching patterns in the library for the measurements that are detected to have interference.

A suite of Perl programs were written to simulate the effect of different types of interference on the isotope distribution of three *E. coli* peptides in the 800-3000 Da mass range, which are recognized as base peptides for the simulation. For each base peptide, 4 charges (1+, 2+, 3+, and 4+) were considered and the intensities based on the natural isotopic abundance of ions were calculated. For each mass and charge combination, the different types of peptide interferences were enumerated. The interference could be in any charge state, have a mass offset (+1, +2, +3, 0, -1, -2, and -3), and different relative intensities (0.1, 0.33, 1.0, 3.0 and 1.0).

Figure 2 shows examples of the interference simulation based on the 2+ charged peptide TKPHVNVGTIGHVHDHGKTT which has 1997 Da mass. In those examples, the charge state of the peptides which causes interference is 4+ and the mass offset is 1.0. The first column shows the intensities of theoretical isotope distribution for TKPHVNVGTIGHVHDHGKTT and the intensities of the monoisotopic peaks are all set to 1.0. The second column shows the average intensities of theoretical isotope distribution for *S. cerevisiae* peptides with mass as 3994 Da, which have same m/z value as 2+ charged TKPHVNVGTIGHVHDHGKTT when they are 4+ charged. From the top to the bottom in the second column, the intensities of the monoisotopic peak in each graph are 0.1, 0.33, 1.0, 3.0, and 0.1 respectively, which covers a range of relative intensities of the interference. The first column and second column overlapped to generate the third column, in which the blue is the correct intensity based on theoretical isotope distribution and the red on the top of the blue is the increased intensity caused by interference. The fourth column shows the increase of intensities for the first three peaks of the theoretical isotope distribution.

When a measurement is detected to have interference, it is searched in the library for matching patterns. Each graph as those in the third column of Figure 2 can be candidate patterns for searching. In order to narrow the searching range, we only focus on the base peptides whose charges are the same as the measured peptide and m/z values within the range of 200 of the measured peptides m/z value. We used Equation 2 to check the similarities between the measured intensities of a peptide and a pattern in the library. To find a matching pattern in the library, the following two conditions must be satisfied:

1. The measured intensities should be similar to the pattern in the library, which means $d_i f_1$ and $d_i f_2$ are both smaller than the threshold 0.1.

2. Without adding interference, the theoretical isotope distribution of the base peptide should be similar to the theoretical isotope distribution of the measured peptide and the threshold is set to 0.1 as well.

3 Results and Discussions

The measurements for 2+ charged GLPADVVPGDILLDDGR in *E. coli* plus lower concentration of *S. cerevisiae* sample (group 2), and *E. coli* plus higher concentration of *S. cerevisiae* sample (group 3) were detected to have interference as mentioned in Section 2.2. After searching in the library, two patterns were found to match with the measurements from group 2 and one pattern was found to match with the measurements from group 3. Figure 3 shows a matching pattern from the library for group 2
Figure 2. Interference simulation examples for 2+ charged TKPHVNVTIGHVDHGKTT. The charge of the peptides which causes interference is 4. For the first three columns, the x axis is the m/z and y axis is the intensity. For the last column, the x axis is the m/z and y axis is the increase of intensity for the first three peaks of theoretical isotope distribution.
Figure 3. Pattern searching for 2+ charged GLPADVVPGDILLDDGR measurement from group 2. (A) The first three peaks of the isotope distribution of GLPADVVPGDILLDDGR. (B) The measured isotope distribution of GLPADVVPGDILLDDGR from group 2. (C) The isotope distribution of the base peptide. (D) The isotope distribution of the peptides that cause interference. (E) The pattern that matches with (B). (F) The increase of intensity from interference for the first three peaks.

measurements. In this pattern the mass of the interference is 1997 Da; the mass offset is -3 and the charge is 2+. Figure 4(A) shows the correction based on the pattern in Figure 3. The corrected intensity is very close to the measured intensity for group 1. Figure 4(B) shows the interference correction for group 2 measurements based on the other pattern, and Figure 4(C) shows the correction for the measurements from group 3.

The interference corrections presented in Figure 4 all reduce the impact of interference for the measurements of 2+ charged *E. coli* peptide GLPADVVPGDILLDDGR. For the two matching patterns of group 2 measurements, the correction showed in Figure 4(B) is not as good as the correction showed in Figure 4(A). However, the pattern matching can be more specific by adjusting the parameters. When we compare the difference between two isotope distributions, how the threshold is set will have influence on the interference detection and correction. If the threshold is increased, less interference will be detected and more patterns will be selected from the library for interference detection. If the threshold is decreased, more interference will be detected and fewer patterns will be selected from the library for interference detection. For the two patterns selected for group 2 measurements mentioned above, if we set the threshold for pattern searching to be 0.9 instead of 1.0, the pattern that was used to generate the correction in Figure 4(B) would not be selected and Figure 4(A) would be the only correction result for the group 2 measurements.
Figure 4. Interference correction result for 2+ charged GLPADVPGDILLDDGR measurement. (A) and (B) are the corrections for group 2, and (C) is the correction for group 3. X axis shows the three groups of data: uncorrected measurement, corrected measurement and measurement from group 1. Y axis is the intensity.

References


