SHORT COMMUNICATION

Sequence Scrambling in Shotgun Proteomics is Negligible

Anton A. Goloborodko,¹ Mikhail V. Gorshkov,¹ David M. Good,² Roman A. Zubarev^{2,3}

¹Institute for Energy Problems of Chemical Physics, Russian Academy of Sciences, Moscow, Russia

²Division of Physiological Chemistry I, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Scheelesväg 2, S-171 77, Stockholm, Sweden

³Science for Life Laboratory, Stockholm, Sweden

Abstract

Analysis of 15,897 low-energy (CAD) and 10,878 higher-energy (HCD) collisional dissociation mass spectra of doubly protonated tryptic peptides taken with high resolution revealed that the rate of sequence scrambling due to b-ion cyclization is negligible (<1%) and can be safely ignored as a possible source of erroneous sequence assignment in shotgun proteomics. On the other hand, there is significant presence of normal (non-scrambled) internal fragments in HCD, which should be taken into account by MS/MS search engines.

Key words: Peptide fragmentation, CAD, HCD, Proteomics, De novo sequencing, Database search

Introduction

Shotgun, or bottom-up, approach is a dominant way to study proteomes with mass spectrometry (MS) [1]. In this approach, proteins are digested, typically by trypsin, and the resultant mixture of tryptic peptides is separated by liquid chromatography coupled on-line with tandem mass spectrometry (MS/MS). Peptide sequences are identified by matching the MS/MS datasets against theoretical spectra of in silico digested proteins from a sequence database [2]. The success of such identification depends upon the quality of MS/MS data, in particular the presence of abundant sequence-specific ions and the absence of interferences. One such interference can arise due to internal fragments that result from cleavage in the precursor ion of more than one peptide bond [3]. Moreover, N-terminal peptide fragments (b ions) have been reported to undergo cyclization followed by secondary backbone cleavage that can open the cycle at a different position than the cyclized peptide bond [4]. Such an opening produces internal backbone fragments of b-type with scrambled amino acid sequences compared with the original peptide. For example, peptide ABCDE can give an ABCD b-ion that could cyclize and, upon ring opening between B and C, produces a fragment CDAB. The

latter fragment can dissociate further, yielding CDA, CD, DAB, and AB, of which CDA and DAB have unexpected masses, while CD is a “normal” internal fragment. Polfer et al. [5] have found cyclization to be dominant process for larger polyglycine b ions containing more than eight residues. Such a process, if frequent, could jeopardize the accuracy of peptide sequence identification in shotgun proteomics.

There is an uncertainty on the effect of b-ion scrambling on sequence assignment in the course of large-scale proteomic experiment [5–7]. The early attempts to evaluate its significance have been limited to small peptide datasets. The purpose of this work is to study the appearance frequency of scrambled fragments through statistical evaluation of this effect in extensive, high-resolution MS/MS datasets. One of these datasets is SwedCAD, a library of 15,897 annotated high-resolution low-energy CAD MS/MS spectra of doubly charged peptides [8]. The second database, SwedHCD, is introduced in this study. It contains 10,878 annotated high-resolution higher-energy (HCD) MS/MS spectra of unique doubly protonated peptides from *Arabidopsis thaliana*, rat, and human samples.

Experimental

Proteomics samples were prepared according to commonly used protocol. Cells were lysed for 10 min at 40 °C in 0.1%

Correspondence to: Roman Zubarev; e-mail: roman.zubarev@ki.se

Figure 1. Number of matches of theoretical m/z values and those experimentally obtained in low-energy CAD MS/MS for: **(a)** conventional backbone cleavage; **(b)** normal internal fragments; **(c)** fragments of b -ions with scrambled sequences; **(d)** fragments of y -ions with scrambled sequences. The horizontal dot-and-dash line shows the estimated number of spurious assignments in each window

solution of Rapigest (Waters). The extracted proteins were reduced for 10 min at 95 °C in 5 mM DTT (Sigma) and alkylated with 10 mM iodoacetamide (Sigma) for 30 min at room temperature. Subsequently, proteins were digested overnight using sequencing-grade trypsin in a 1:50 trypsin:protein ratio. The tryptic peptides were filtered using a 10 kDa cut-off filter (Pall Life Science) and dried in a speed vac.

Peptide mixtures were loaded into either HP1100 (Agilent) or EASY-nLC (Proxeon) equipped with a home-packed, sprayer-integrated 15 cm nano-column (Proxeon). Mass spectrometric analysis with low-energy CAD was carried out using a hybrid LTQ-Orbitrap [9, 10], and Orbitrap Velos (both - ThermoFisher Scientific) for higher-energy dissociation. Both MS and MS/MS ion detection was performed in an FT analyzer at >10,000 resolving power.

Resultant .raw files were converted into Mascot Generic Format (.mgf) files using DTA Generator [11, 12]. Data were searched against a concatenated forward and reversed databases using MASCOT search engine (Matrix Science). All resulting sequences were trimmed to a <1% false discovery rate [13] before inclusion into the SwedCAD or SwedHCD database. Redundant sequences were removed with the highest scoring peptides retained.

For each fragmentation spectrum in SwedCAD and SwedHCD databases, theoretical m/z values of fragments were generated and matched with 10 ppm tolerance against the experimentally obtained m/z values that were preliminarily de-charged and de-isotoped. Each experimental m/z

value could be attributed only to a single fragment type in the following order: conventional backbone fragments (a , b , c , x , y , and z with possible loss of H_2O and NH_3 from a and b); internal fragments from the original sequence; fragments from the scrambled sequences. To monitor the number of spurious matches, y -ion sequences were scrambled. Since cyclization of y -ions has not been reported and it is highly unlikely from the mechanistic point of view, while y -ions have the same amino acid composition as b -ions, scrambled y -ions provide excellent decoy dataset for scrambled b -ions. As an additional control, theoretical m/z values of scrambled b -ions were shifted by +3 Th.

Results

While, as expected, conventional b , y fragments comprised a major part of the identified ions in the low-energy SwedCAD data (173,731 matches, Figure 1a), internal fragments gave a much smaller number of assignments (3215 ions or ~2%, Figure 1b). The remaining ions gave close numbers of hits when matched against fragments of scrambled b - (4717 hits, Figure 1c) and y -ions (4880 hits, Figure 1d) or even 3 Th shifted scrambled b -ions (5691 hits; data not shown).

For the high-energy data, the frequency of internal fragments among identified ions was much higher and comprised 18% of that for conventional fragments (12,919 versus 75,220 hits, Figure 2b and a, respectively). At the same time, the frequency of matched fragments of scrambled

Figure 2. Number of matches of theoretical m/z values and those experimentally obtained in higher-energy HCD MS/MS for: **(a)** conventional backbone cleavage; **(b)** normal internal fragments; **(c)** fragments of b -ions with scrambled sequences; **(d)** fragments of y -ions with scrambled sequences. The horizontal dot-and-dash line shows the estimated number of spurious assignments in each window

b -ions, y -ions, and 3 Th shifted scrambled b -ions remained between 2.5% and 3.0%, 2209 (Figure 2c), 2474 (Figure 2d) and 2153 hits (not shown), respectively.

Since matches to scrambled y -ions and mass shifted scrambled b -ions are of spurious nature, and the number of such matches is similar to that of matches to scrambled b -ions, the formation rate of the latter species is within the statistical uncertainty caused by false positive discovery. To estimate the number of definitely wrong assignments among the detected hits, we assumed that the spurious assignments are distributed uniformly over the mass window. By subtracting from the total hit number the areas of the pedestals in the mass error histograms in Figures 1 and 2, we estimated the upper limit for true positive hits. For CAD MS/MS spectra, internal fragments gave maximum 1938 hits (1.3% of the normal b and y hits), scrambled fragments of b -ions, 1381 hits (<1 %), y -ions, 1458 hits, and 3 Th shifted scrambled fragments of b -ions, 1001 hits. For HCD MS/MS spectra, maximum 11,204 internal fragments were detected, compared with 1037 hits for scrambled b -ion, 1008 for scrambled y -ions, and 338 for 3 Th shifted b -ion fragments. Assuming Poisson statistics, standard deviation of a hit number is equal to the square root of that number. Thus, the number of matches for the scrambled b - and y -ions are the same within less than two standard deviations in both CAD and HCD, rendering them statistically indistinguishable. If we consider the second control, +3 Th shifted scrambled b -ions, the enhanced occurrence of scrambled b -ions becomes

statistically significant, but still much less than 1% of the total number of matched fragments.

Note that the majority of previous studies demonstrating scrambling were performed on singly-charged, non-tryptic, and rather small peptides [4, 5, 7, 14]. We believe that the different outcomes in these studies compared with the current one were determined by the differences in the peptide nature and especially the charge state and the sequence length.

Conclusions

The rate of b -ion cyclization in real-life shotgun proteomics is negligible (<1%), and can be safely ignored as a possible source of error in sequence assignment. On the other hand, there is significant presence of normal internal fragments in HCD, which should be taken into account by MS/MS search engines.

Acknowledgment

The authors acknowledge support for this work by the Swedish Research Council (grant 2007–4410 to R.A.Z.). D. M.G. received a post-doctoral fellowship from the Wenner-Gren Foundation. M.V.G. thanks Russian Foundation for Basic Research, grant no. 11-04-00515-a

References

1. Aebersold, R., Mann, M.: Mass Spectrometry-Based Proteomics. *Nature* **422**, 198–207 (2003)
2. Link, A.J., Eng, J., Schieltz, D.M., Carmack, E., Mize, G.J., Morris, D. R., Garvik, B.M., Yates, J.R.: Direct Analysis of Protein Complexes Using Mass Spectrometry. *Nat. Biotechnol.* **17**, 676–682 (1999)
3. Paizs, B., Suhai, S.: Fragmentation Pathways of Protonated Peptides. *Mass Spectrom. Rev.* **24**, 508–548 (2005)
4. Harrison, A.G., Young, A.B., Bleiholder, C., Suhai, S., Paizs, B.: Scrambling of Sequence Information in Collision-Induced Dissociation of Peptides. *J. Am. Chem. Soc.* **128**, 10364–10365 (2006)
5. Chen, X., Yu, L., Steill, J.D., Oomens, J., Polfer, N.C.: Effect of Peptide Fragment Size on the Propensity of Cyclization in Collision-Induced Dissociation: Oligoglycine b2–b8. *J. Am. Chem. Soc.* **131**, 18272–18282 (2009)
6. Harrison, A.G.: To b or Not to b: The Ongoing Saga of Peptide b Ions. *Mass Spectrom. Rev.* **28**, 640–654 (2009)
7. Molesworth, S., Osburn, S., Van Stipdonk, M.: Influence of Size on Apparent Scrambling of Sequence During CID of b-Type Ions. *J. Am. Soc. Mass Spectrom.* **20**, 2174–2181 (2009)
8. Fälfth, M., Savitski, M.M., Nielsen, M.L., Kjeldsen, F., Andren, P.E., Zubarev, R.A.: SwedCAD, a Database of Annotated High-Mass Accuracy MS/MS Spectra of Tryptic Peptides. *J. Proteome Res.* **6**, 4063–4067 (2007)
9. McAlister, G.C., Phanstiel, D., Good, D.M., Berggren, W.T., Coon, J.J.: Implementation of Electron-Transfer Dissociation on a Hybrid Linear Ion Trap-Orbitrap Mass Spectrometer. *Anal. Chem.* **79**, 3525–3534 (2007)
10. McAlister, G.C., Berggren, W.T., Griep-Raming, J., Horning, S., Makarov, A., Phanstiel, D., Stafford, G., Swaney, D.L., Syka, J.E.P., Zabrouskov, V., Coon, J.J.: A Proteomics Grade Electron Transfer Dissociation-Enabled Hybrid Linear Ion Trap-Orbitrap Mass Spectrometer. *J. Protein Res.* **7**, 3127–3136 (2008)
11. Good, D.M., Wenger, C.D., Coon, J.J.: The Effect of Interfering Ions on Search Algorithm Performance for Electron-Transfer Dissociation Data. *Proteomics* **10**, 164–167 (2010)
12. Good, D.M., Wenger, C.D., McAlister, G.C., Bai, D.L., Hunt, D.F., Coon, J.J.: Post-Acquisition ETD Spectral Processing for Increased Peptide Identifications. *J. Am. Soc. Mass Spectrom.* **20**, 1435–1440 (2009)
13. Elias, J.E., Gygi, S.P.: Target-Decoy Search Strategy for Increased Confidence in Large-Scale Protein Identifications by Mass Spectrometry. *Nat. Methods* **4**, 207–214 (2007)
14. Harrison, A.G.: Peptide Sequence Scrambling Through Cyclization of b (5) Ions. *J. Am. Soc. Mass Spectrom.* **19**, 1776–1780 (2008)