# Protein de Novo Sequencing by Top-Down and Middle-Down MS/MS: Limitations Imposed by Mass Measurement Accuracy and Gaps in Sequence Coverage

He, L. (FSU, Chemistry); Weisbrod, C.R. (NHMFL, ICR) and Marshall, A.G. (FSU, Chemistry; NHMFL, ICR)

## Results and Discussion

With the increasing accessibility of Fourier transform (FT) mass spectrometers, top-down/middle-down MS/MS characterization of protein sequences is rapidly gaining popularity. Compared to conventional bottom-up sequencing, the top-down/middle-down approach offers the advantages of fast sample preparation and unambiguous characterization of proteoforms in a mixture. If the modified or mutated peptide segment of interest is not found or recovered with the bottom-up approaches, top-down becomes attractive relative to spending time seeking the right enzyme or chromatographic approach. Here, we discuss the potential and limitations of protein sequence analysis by top-down/middle-down MS/MS alone. Even if 100% protein sequence coverage is achieved by MS/MS, fragment mass error tolerance as low as 1 ppm or 0.5 ppm is needed to differentiate glutamine from lysine at positions not exceeding 330 amino acids (AAs) or 660 AAs from the N-/C-terminus for a protein with 660 AAs (72,760 Da) or 1320 AAs (145,520 Da). To characterize the "AA sequence gap" between two adjacent fragments, we show that the number of gap AA sequences with identical masses for di-, tri-, and tetra-AA gaps grows exponentially with increasing number of gap amino acids. If peptide fragment mass could be measured exactly (in practice, to 0.00001 Da), it would then be possible to define the overall atomic composition for the group of amino acids spanning a product ion gap 3-4 amino acids long. However, when we consider any 3-4 amino acid gap, we find that 50-75% of the possible compositions describe at least two sets of amino acids (see Figure 1). Moreover, a next-generation protein fragment deconvolution algorithm is critical to exploit the experimentally observed high mass accuracy generated from the 21 T FT-ICR MS/MS for high confidence and high throughput top-down/middle-down analysis of proteins with unknown sequences. Finally, we show that de novo top-down/middle-down MS/MS can determine the germline sequence category for a given monoclonal antibody (mAb) and further serve to identify its novel mutations.

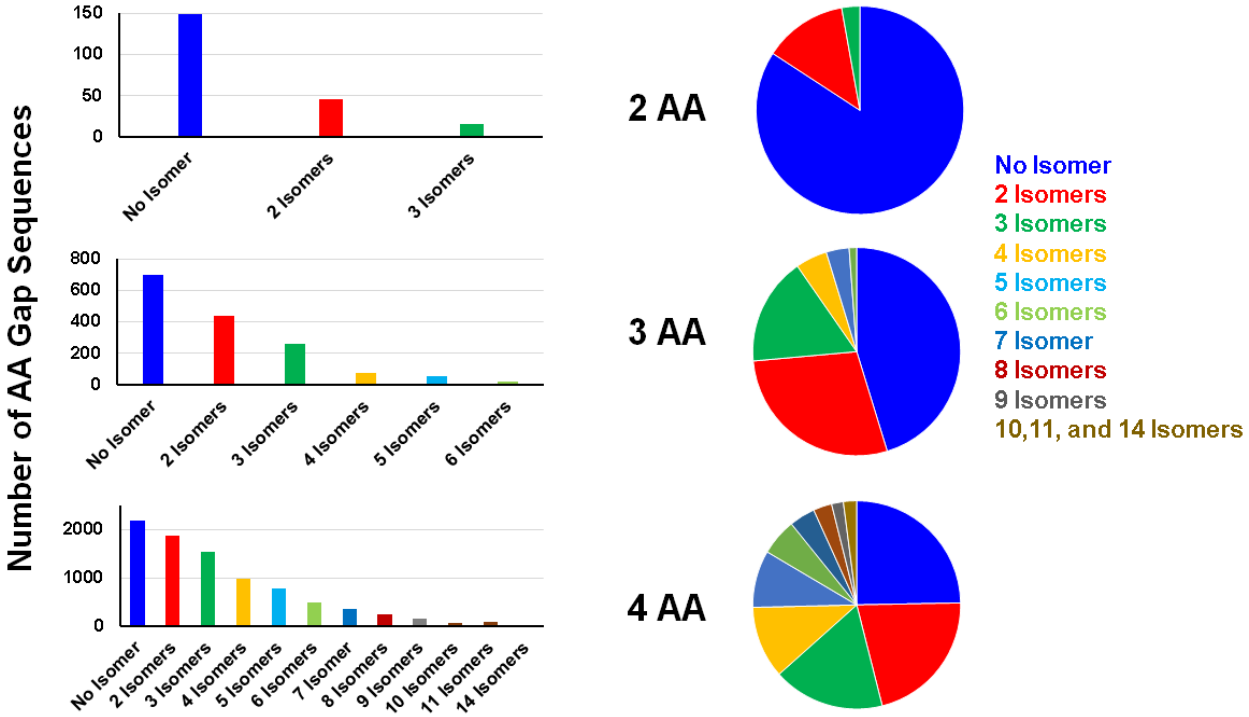


**Fig. 1.** Left: Number of amino acid compositions of the same exact mass, sorted according to the number of isomers for each composition. Right: Percentage for each of the categories shown at the left. Note that even with perfect mass measurement accuracy, more than half of the possible tri-AA gap compositions have at least 2 isomers, and ~75% of the possible tetra-AA gap compositions have at least 2 isomers.

## Acknowledgments

The authors thank Christopher L. Hendrickson and Lissa C. Anderson for helpful discussions. This work was supported by NSF Division of Materials Research (DMR-11-57490) and the State of Florida.

## References

[1] He, L.; *et al.*., Int. J. Mass Spectrom, **427**, 107-113 (2018).