Journal of proteome research

pubs.acs.org/jpr

Technical Note

Theoretical Considerations for Next-Generation Proteomics

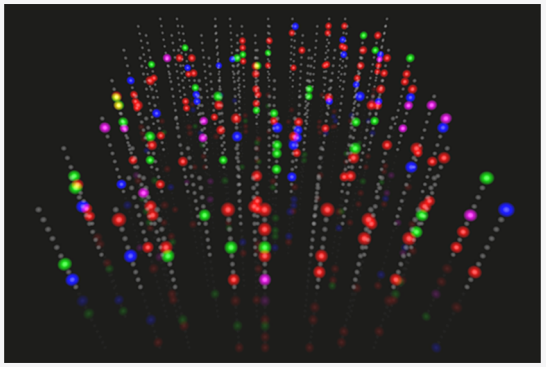
Magnus Palmblad*

Cite This: *J. Proteome Res.* 2021, 20, 3395−3399

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: While mass spectrometry still dominates proteomics research, alternative and potentially disruptive, next-generation technologies are receiving increased investment and attention. Most of these technologies aim at the sequencing of single peptide or protein molecules, typically labeling or otherwise distinguishing a subset of the proteinogenic amino acids. This note considers some theoretical aspects of these future technologies from a bottom-up proteomics viewpoint, including the ability to uniquely identify human proteins as a function of which and how many amino acids can be read, enzymatic efficiency, and the maximum read length. This is done through simulations under ideal and non-ideal conditions to set benchmarks for what may be achievable with future single-molecule sequencing technology. The simulations reveal, among other observations, that the best choice of reading N amino acids performs similarly to the average choice of N+1 amino acids, and that the discrimination power of the amino acids scales with their frequency in the proteome. The simulations are agnostic with respect to the next-generation proteomics platform, and the results and conclusions should therefore be applicable to any single-molecule partial peptide sequencing technology.



KEYWORDS: *simulation, theory, R, next-generation proteomics, single-molecule sequencing, fluorosequencing, enzymatic digestion, peptide−partial read match, protein identification, NeXtProt*

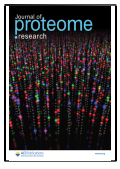
INTRODUCTION

For three decades, mass spectrometry has dominated the field of proteomics and has been the primary method for protein identification, characterization, and quantitation. Over these years, mass spectrometry has seen tremendous growth in speed and sensitivity. Fundamentally, however, it remains a serial analysis technique. Recent developments and simulations of single-molecule and massively parallel detection techniques suggest a disruptive technological breakthrough may be closer than many in the field realize. It is therefore worth considering what these breakthroughs may look like through the lens that we have—from the state-of-the-art in computational and mass spectrometry-based proteomics.

Callahan et al.[1] and Winston and Gregory Timp[2] recently surveyed the technologies most promising to disrupt the technological status quo. Some of these are direct analogs of nucleic acid sequencing technologies, such as fluorosequencing[3] common in next-generation sequencing, and nanopores for protein[4] similar to those used for reading long stretches of DNA. In addition, there are several antibody-based methods that not only more or less specifically detect proteins with high sensitivity but also localize them in cells or tissues, including CITE-Seq[5] for concurrent mRNA and protein detection, as well as the possibility of reverse translation coupled with DNA sequencing readout.[6] Both fluorosequencing[7,8] and nanopore protein sequencing[9] have been simulated to show that the partial reads they generate can uniquely identify proteins in a given proteome, analogous to how tandem mass spectra are matched to peptides in bottom-up proteomics. The analogy is not perfect, as in mass spectrometry different peptides produce different tandem mass spectra, even if they only differ by a leucine in place of an isoleucine,[10] unlike partial fluorosequencing, which does not distinguish between unlabeled residues.

This paper explores the potential—and possible limitations—of the next generation of massively parallel proteomic technologies (Figure 1), specifically investigating the relationships between the number of readable amino acid residues, read length, and ability to unambiguously identify human proteins. Many other variables are likely to influence protein detection limits and eventual acceptance of a different technology in proteomics, but these are beyond the scope of this short paper. Focusing on short tryptic peptides as common in bottom-up proteomics, we simulate hypothetical single-molecule sequencing experiments under ideal and non-ideal conditions, where the former represent a possible best performance and the latter what

Received: February 17, 2021
Published: April 27, 2021



https://doi.org/10.1021/acs.jproteome.1c00136
J. Proteome Res. 2021, 20, 3395−3399

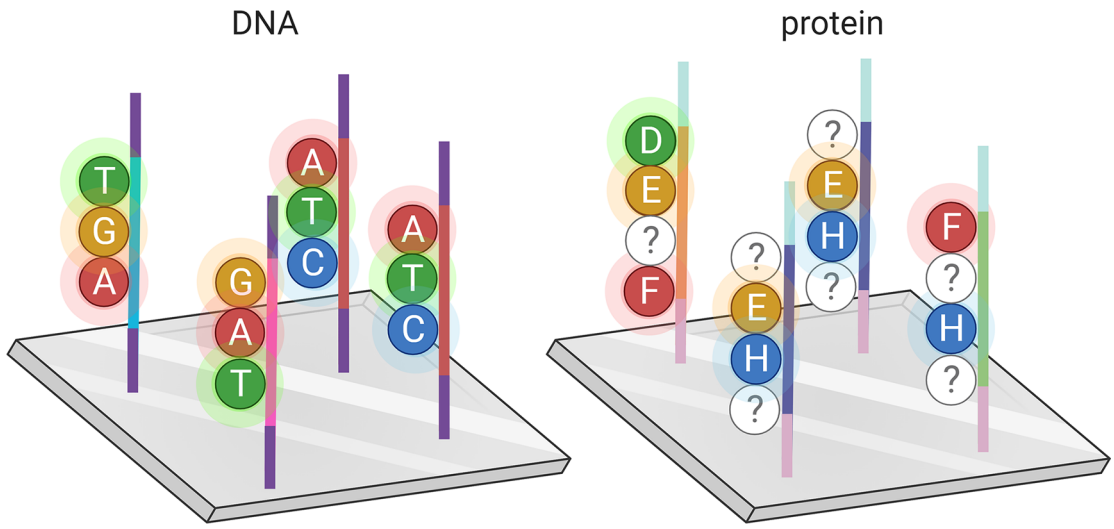


Figure 1. General analogy of next-generation genomics and proteomics, with single-molecule sequencing of some of the 20 amino acids. Unlike DNA sequencing, where oligonucleotides can be amplified on the chip, peptide sequencing requires true single-molecule sensitivity. Each peptide is individually identified by matching the partial read to peptides derived from a sequence database, producing a peptide−partial read match (PPRM). Quantification can be analogous by counting reads, similar to counting spectra, or peptide−spectrum matches (PSMs) in mass spectrometry-based proteomics. This is also how transcripts are quantified in RNA-Seq experiments. With several orders of magnitude more PPRMs in one experiment than PSMs in a typical LC-MS/MS run, next-generation proteomics will have a superior dynamic range.

we could realistically expect with imperfect technology. To be as general as possible, we will not make specific assumptions on the experimental technology, but for the sake of these thought experiments we assume the technology will be able to read perfectly the visible amino acids up to a certain read length and that there is no interference between reads caused by similar sequences, post-translational modifications, or relative abundance.

METHODS

To simulate the coverage of the human proteome by short partial reads, the experiment was simulated by an R script. The script fetches all NeXtProt sequences (currently 20 322) from UniProt using the *UniProt.ws* package 2.28.0,[11] and digests these using the *cleaver* package version 1.26.1.[12] Sequence variants and post-translational modifications are ignored. After blanking out the invisible amino acids, the proteotypic peptide−partial read matches (PPRMs) are tallied for each protein, resulting in a number of proteins without unique PPRMs. How this number is influenced by the composition and frequency of the labeled amino acids, how long reads are generated, and the enzymatic efficiency were subsequently simulated. Each combination of variables was simulated 20 times. The R code for the simulations, the simulation results, and the R script for making the plots in Figure 2 are all available in GitHub.[13]

To estimate the power of the different amino acids to discriminate between (human) proteins, an additional 100 random combinations of 4, 5, and 6 readable amino acids were simulated and added to the 20 combinations already simulated, for a total of 360 simulations (out of 59 109 possibilities). The resulting number of uniquely identifiable proteins were fitted using linear regression in R, and the coefficients for each amino acid compared, similar to the first-generation retention time prediction.[14]

Non-ideal conditions were simulated by allowing for one missed invisible amino acid in each gap in the read sequence, recalculating the number of proteins with unique proteotypic reads given the additional possible peptide reads.

RESULTS AND DISCUSSION

Only reading one or two amino acids is unlikely to be sufficient to disrupt the current technological dominance of mass spectrometry. However, reading as few as three amino acids, it is already possible to distinguish over 19 800, or 97.5% of the proteins in NeXtProt, when allowing for one missed cleavage and read length up to 50 (Figure 2A,B, compare with Figure 2 in Swaminathan et al.[7]). Some readers may remember this is more than the number of proteins quantifiable by the first-generation commercial isotopic reagents for proteomics[15] that were unable to label the nearly 600 human proteins lacking cysteines. This limitation did not prevent these reagents from being successfully adopted by the proteomics community, eventually leading to improved labeling schemes such as TMT.[16] There is little reason to assume that being unable to uniquely quantify this number of proteins would be an unforgivable shortcoming. This degree of coverage can be achieved with as few as three readable amino acids if allowing reads up to 50 residues, and seven readable amino acids if limited to reads up to 20 residues in length. It should also be noted that even *if* we can read all 20 amino acids, 227 proteins in NeXtProt still lack unique tryptic peptides with zero or one missed cleavage ≤50 amino acids in length (Figure 2A,B). Allowing for a missed cleavage site, which is commonly done for peptide identification in bottom-up proteomics, makes only a small difference (Figure 2B), though some longer proteotypic reads can be matched when allowing for missed cleavages. Trivially, all simulations converge at the reading of 19 out of the 20 amino acids (Figure 2A,B), as the single invisible amino acid can always be inferred.

Shorter maximum read lengths require more amino acids to be visible for unambiguous protein identification. With a maximum read length of 10, one needs to read at least 10 amino acids for 95% of the proteins in NeXtProt to have a proteotypic read (Figure 2A,C). Trypsin does not generate many peptides longer than 50 amino acids from proteins digested within one missed cleavage from completion, and there is no apparent benefit for protein identification in running more cycles to generate a few longer reads, even if this would prove

https://doi.org/10.1021/acs.jproteome.1c00136
J. Proteome Res. 2021, 20, 3395−3399

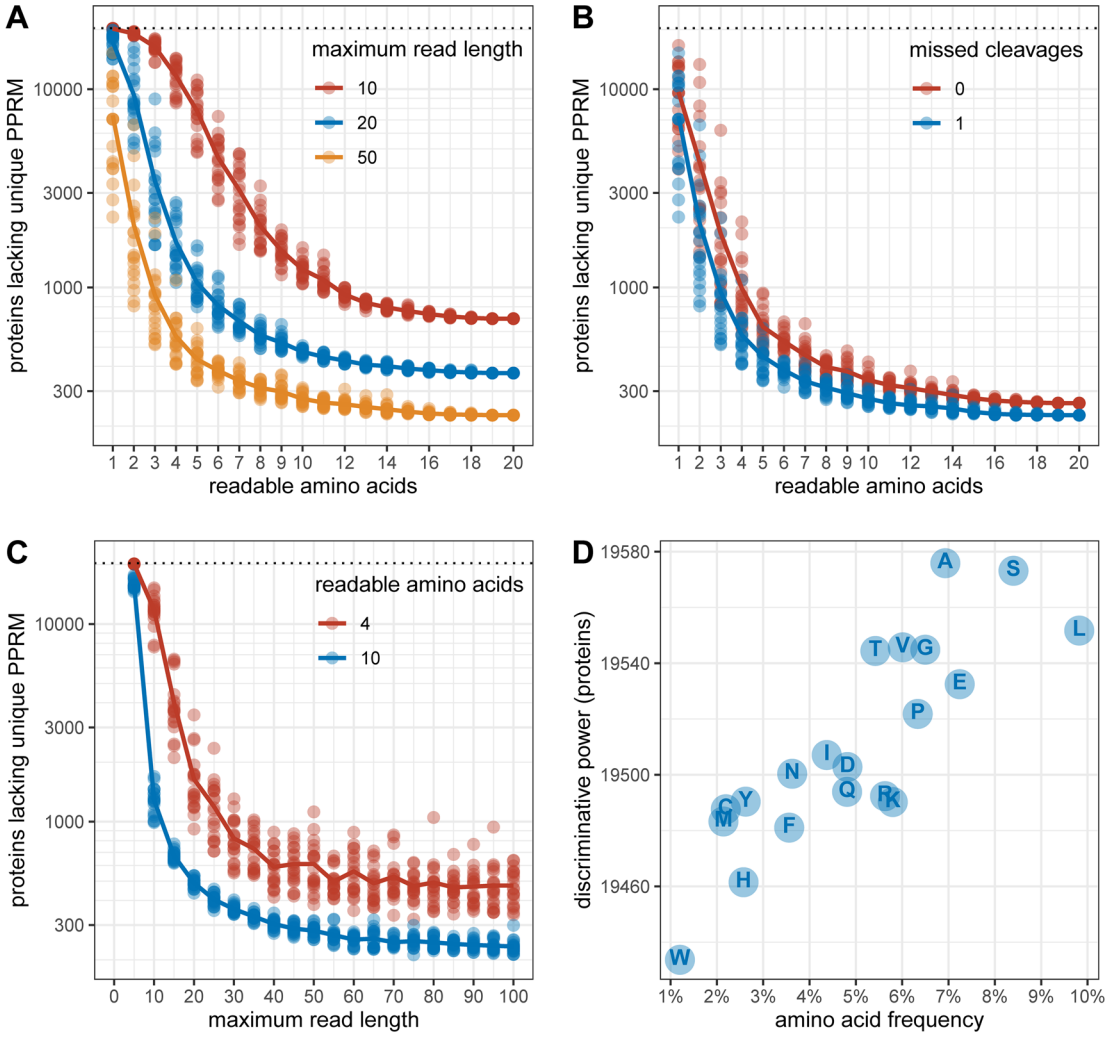


Figure 2. Simulation results showing the number of proteins lacking a unique PPRM as a function of the number of readable, or "visible", amino acids, comparing maximum read lengths of 10, 20, and 50 with one missed cleavage allowed (A) and zero or one missed cleavage with read length ≤50 (B). Panel C shows the number of proteins without unique reads as a function of the maximum read length for 4 and 10 readable amino acids. Panel D displays the discriminative power of the amino acids when reading 4−6 amino acids (read length ≤50) plotted against frequency in NeXtProt. The data for read length ≤50 and one missed cleavage in A (yellow) and B (blue) are the same. The dotted lines indicate the 20 322 proteins in the current version of NeXtProt.

possible (Figure 2C). However, this observation may not hold for other proteases, and there is no guarantee trypsin will remain the enzyme of choice for technologies very different from LC-MS/MS. When labeling/reading more amino acids, which ones are read is apparently less important (Figure 2C). On average, the more frequent amino acids contribute more to the unambiguous protein identification when reading between 4 and 6 different amino acids (Figure 2D). This is true in each case, though the slope and intersect depend on the number of read amino acids. In the extreme case of reading only one amino acid, serine, alanine, and leucine are also the best choices under the metric used here, with tryptophan the worst. Knowing that a peptide contains alanine is less informative than knowing it contains a tryptophan. However, a partial sequence with several alanines (e.g., ---A----A--A---A--) can be more informative than a partial sequence with a single tryptophan (e.g., -----W----------). On average we observe many more alanines, and therefore partial sequences with multiple alanines, than we see tryptophans per peptide. Minor possible outliers are cysteine (C) and methionine (M), being more informative than other some more frequent amino acids, and arginine (R) and lysine (K) being less discriminating between proteins. The latter may in part be due to these residues defining the cleavage site, and as only one missed cleavage site was allowed, no read has more than two of these residues in total, including the peptide C-terminal residue. The improvements by reading more than six different amino acids when allowing for longer reads and missed cleavages are very small and only noticeable in Figure 2 due to the logarithmic scale of the ordinate. However, this compares with mass spectrometry-based proteomics as currently performed by most practitioners in the field. To be able to robustly analyze single amino acid variants or semi-tryptic and non-tryptic peptides, or to distinguish proteins from different species in metaproteomes,[17] there will likely be significant benefit from reading a few additional amino acids. It cannot be ruled out with certainty that the optimum combination of N amino acids is surrounded by poor ones. Finding the absolute best solution therefore requires brute-force calculation of all possibilities. The worst case is considering 10 out of 20 amino acids, with "20 choose 10" or 184 756 combinations. This is well within the

https://doi.org/10.1021/acs.jproteome.1c00136
J. Proteome Res. 2021, 20, 3395−3399

reach of high-performance computing, but here we will limit ourselves to looking at the average performance at a given number of readable amino acids, as this still reveals any general trends. Looking at Figure 2A, we also note that the best performance of reading N amino acids is similar to the average performance of reading N+1 amino acids for all curves. This and the other key observations from the simulations are summarized as simplified "rules" in Table 1.

Table 1. "Rules" Summarizing the Three Main Observations Made from the Simulations in This Work

Rule 1:	The *best* selection of N amino acids provides a similar number of proteotypic reads as the *average* selection of N+1 amino acids.	Figure 2A,B
Rule 2:	There is no benefit in reading beyond 50 residues of tryptic peptides.	Figure 2C
Rule 3:	The discriminatory power of the amino acids scales with their frequency in the proteome.	Figure 2D

The simulations of the non-ideal experiment, where some invisible amino acids are not cleaved (or moved through a pore, or reverse translated) and therefore randomly extend some gaps, revealed that the number of proteins without proteotypic partial reads is only moderately affected, and only when a small number of amino acids are visible (Figure 3), even though the number of possible reads increased approximately 6-fold (on average from 0.8 to 4.5 million when reading 4 amino acids, from 1.4 million to 8.6 million when reading 10). This demonstrates that informative reads can still be generated with sub-optimal yield in the sequencing steps.

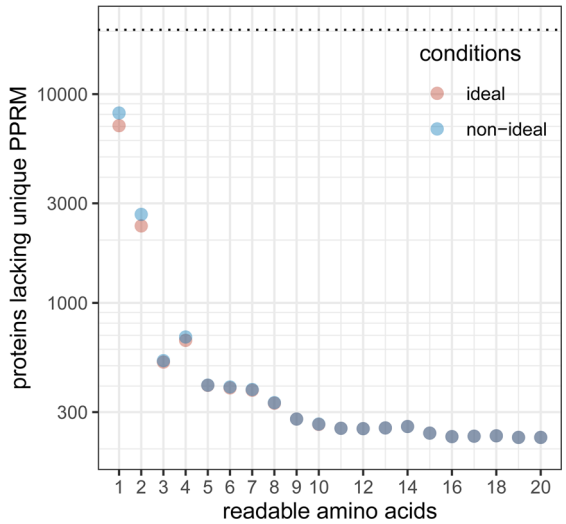


Figure 3. Results from simulation under non-ideal conditions, with read length ≤50 and one missed tryptic cleavage where in each sequence gap one step may be missed, resulting in a longer gap. For example, reading only alanines in a peptide WANDA results in not only the read -A--A, but also --A--A, -A---A, and --A---A. Though the number of possible reads is higher (approximately 6-fold when reading 4 or 10 amino acids), the number of proteins lacking unique PPRMs is only moderately affected (here showing the exact same selections of 1−20 amino acids under ideal and non-ideal conditions for clarity). The dotted line indicates the 20 322 proteins in NeXtProt.

Here we used as target metric the number of proteins in NeXtProt lacking a proteotypic partial read of a tryptic peptide. The optimum choice of amino acids to read, given this metric and assuming one has a choice, depends on the proteome and organism studied. This is just one of many possible targets. One could instead opt to maximize the number of PPRMs to get the most quantitative information out of an experiment. The number of read cycles (maximum read length) or yields may also be more constrained. Traditional Edman sequencing, for example, struggles with peptides much longer than 30 amino acids, even though single-molecule sequencing may not suffer the same limitations.

CONCLUSIONS

The next technological disruption in proteomics will likely come from massive parallelism rather than increased sensitivity over mass spectrometry and Edman sequencing (which at best are in the zeptomole[18] and low-attomole[19] ranges), though the increased depth of single-cell proteomics from single-molecule detection could be expected to be dramatic. When next-generation proteomics of a type simulated here and by Swaminathan et al.[7] is realized, we will need algorithms to match the partial reads to peptides that do not make as strict and simplistic assumptions as in the simulations here. For example, it may not be justified to assume we know the number of consecutive invisible residues, or even the length of the read, and the matching has to be tolerant for these and other effects stemming from labeling failures, photobleaching, or sub-optimal yield in the chemical or enzymatic steps during sequencing (or reverse translation). This will undoubtedly be a hot topic for computational proteomics in coming years, and error-tolerance will likely be the "name of the game" in single-molecule protein identification. Though this work is entirely theoretical in nature, combining known characteristics of underlying biology with realistic assumptions of future technology underscores just how close that future is.

AUTHOR INFORMATION

Corresponding Author

Magnus Palmblad − *Center for Proteomics and Metabolomics, Leiden University Medical Center, Leiden 2300 RC, The Netherlands;* orcid.org/0000-0002-5865-8994; Email: n.m.palmblad@lumc.nl

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jproteome.1c00136

Notes

The author declares no competing financial interest.

ACKNOWLEDGMENTS

This paper is dedicated to my late mentor and scientific hero, Professor André M. Deelder (1947−2021). Dr. Benjamin Neely is also thanked for numerous helpful comments and suggestions. Figure 1 was created in BioRender.

REFERENCES

(1) Callahan, N.; Tullman, J.; Kelman, Z.; Marino, J. Strategies for Development of a Next-Generation Protein Sequencing Platform. *Trends Biochem. Sci.* **2020**, *45* (1), 76−89.

(2) Timp, W.; Timp, G. Beyond mass spectrometry, the next step in proteomics. *Sci. Adv.* **2020**, *6* (2), No. eaax8978.

(3) Swaminathan, J.; Boulgakov, A. A.; Hernandez, E. T.; Bardo, A. M.; Bachman, J. L.; Marotta, J.; Johnson, A. M.; Anslyn, E. V.; Marcotte, E. M. Highly parallel single-molecule identification of proteins in zeptomole-scale mixtures. *Nat. Biotechnol.* **2018**, *36*, 1076.

(4) Kennedy, E.; Dong, Z.; Tennant, C.; Timp, G. Reading the primary structure of a protein with 0.07 nm(3) resolution using a subnanometre-diameter pore. *Nat. Nanotechnol.* **2016**, *11* (11), 968−976.
(5) Stoeckius, M.; Hafemeister, C.; Stephenson, W.; Houck-Loomis, B.; Chattopadhyay, P. K.; Swerdlow, H.; Satija, R.; Smibert, P. Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **2017**, *14* (9), 865−868.
(6) Martin, M. T. Methods and compositions for reverse translation. U.S. Patent 7169894 B2, Jan 30, 2007.
(7) Swaminathan, J.; Boulgakov, A. A.; Marcotte, E. M. A theoretical justification for single molecule peptide sequencing. *PLoS Comput. Biol.* **2015**, *11* (2), No. e1004080.
(8) Rodriques, S. G.; Marblestone, A. H.; Boyden, E. S. A theoretical analysis of single molecule protein sequencing via weak binding spectra. *PLoS One* **2019**, *14* (3), No. e0212868.
(9) Ohayon, S.; Girsault, A.; Nasser, M.; Shen-Orr, S.; Meller, A. Simulation of single-protein nanopore sensing shows feasibility for whole-proteome identification. *PLoS Comput. Biol.* **2019**, *15* (5), No. e1007067.
(10) Jiang, C.; Arthur, C. J.; Gates, P. J. A computational and experimental study of the fragmentation of l-leucine, l-isoleucine and l-allo-isoleucine under collision-induced dissociation tandem mass spectrometry. *Analyst* **2020**, *145* (20), 6632−6638.
(11) Carlson, M.; Ortutay, C. *UniProt.ws: R Interface to UniProt Web Services*, version 2.28.0; Bioconductor Package Maintainer, 2020.
(12) Gibb, S. *cleaver: Cleavage of Polypeptide Sequences*, version 1.26.1, 2020. https://github.com/sgibb/cleaver/.
(13) Palmblad, M. *NGP*, version 1.0.0, 2021. https://github.com/magnuspalmblad/NGP.
(14) Palmblad, M.; Ramstrom, M.; Markides, K. E.; Hakansson, P.; Bergquist, J. Prediction of chromatographic retention and protein identification in liquid chromatography/mass spectrometry. *Anal. Chem.* **2002**, *74* (22), 5826−30.
(15) Gygi, S. P.; Rist, B.; Gerber, S. A.; Turecek, F.; Gelb, M. H.; Aebersold, R. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* **1999**, *17* (10), 994−9.
(16) Thompson, A.; Schafer, J.; Kuhn, K.; Kienle, S.; Schwarz, J.; Schmidt, G.; Neumann, T.; Hamon, C. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.* **2003**, *75* (8), 1895−1904.
(17) Saito, M. A.; Dorsk, A.; Post, A. F.; McIlvin, M. R.; Rappe, M. S.; DiTullio, G. R.; Moran, D. M. Needles in the blue sea: sub-species specificity in targeted protein biomarker analyses within the vast oceanic microbial metaproteome. *Proteomics* **2015**, *15* (20), 3521−31.
(18) Belov, M. E.; Gorshkov, M. V.; Udseth, H. R.; Anderson, G. A.; Smith, R. D. Zeptomole-sensitivity electrospray ionization–Fourier transform ion cyclotron resonance mass spectrometry of proteins. *Anal. Chem.* **2000**, *72* (10), 2271−9.
(19) Miyashita, M.; Presley, J. M.; Buchholz, B. A.; Lam, K. S.; Lee, Y. M.; Vogel, J. S.; Hammock, B. D. Attomole level protein sequencing by Edman degradation coupled with accelerator mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98* (8), 4403−8.

https://doi.org/10.1021/acs.jproteome.1c00136
J. Proteome Res. 2021, 20, 3395−3399