


The Importance of Considering Natural Isotopes in Improving Protein Identification Accuracy

Sara El Jadid¹ ^a, Raja Touahni¹ and Ahmed Moussa²

¹Faculty of Science, Ibn Tofail University, Kenitra, Morocco

²National School of Applied Science, Abdelmalek Essadi University, Tangier, Morocco

Keywords: Proteomics, Mass Spectrometry, Peptide Identification, Quantitation, Accuracy, Natural Isotopes.

Abstract: Many tools in proteomics are based on accurate identification of peptide contained in a sample. In fact, the issue of identification is the foundation of the entire proteomics workflow, where all subsequent steps depend on the quality of data generated at the beginning. The accuracy of data generated allow, not only to have good results, but also to ensure consistency at the end of the analysis. There is a consensus about the factors that affect this accuracy. It is popularly assumed that exploiting physics and chemistry of peptides deduced from sequences can improve the identification accuracy. In fact, considering natural isotopes when quantifying peptides will considerably improve results. This paper presents findings that defend such a view. We explored the mass difference between the nominal mass (which considers the most abundant isotope of each element) and the mean mass (which considers the abundance of each element). We noticed that within a biomolecule, the larger the number of elements, the less this difference is negligible. In accordance with that, peptide misidentification is due to the previously explained variance. These findings reveal that including natural isotopes during quantification will play a key role in improving identification accuracy. This study could lead us to design alternative identification tools combining better sensitivity and specificity.

1 INTRODUCTION

Proteins perform a large number of functions within the organisms, including catalyzing metabolic reactions, structural organization, DNA replication, transporting molecules from one location to another and integration of internal and external signals (Pratt et al., 2002). Proteins are the primary mediators and the executive core of the cellular phenotype (Schmidt et al., 2014; Aebersold and Mann, 2003). A protein consists of at least one long chain of amino acid residues (Figure 1).

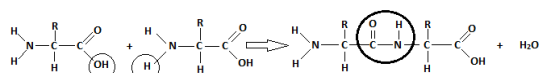



Figure 1: Production of dipeptide from two amino acids. A peptide bond was marked by a big circle.

An amino acid is made up of an amine group, a carboxyl group and a side chain (R group) specific to each of 20 amino acids (Koehler et al., 2011) (Table 1).

^a  <https://orcid.org/0000-0001-9793-5061>

The study of the whole protein set cell is known as proteomics. This downstream "omics" of science is associated with different technologies which allow the separation of proteins (Hunt et al., 1986; Listgarten and Emili, 2005). Over recent years mass spectrometry (MS) made tremendous progress and became the most comprehensive and versatile tool for studying proteins on a large-scale (Mann and Kelleher, 2008; Yates et al., 2009). Mass spectrometry allows to determine with accuracy and sensitivity the mass of molecules (e.g. biomolecules: peptides, proteins, oligonucleotides, sugars, lipids, metabolites) (Deutsch et al., 2010; Gerber et al., 2003). It is a method for measuring the mass-to-charge ratios (m/z : where m is the mass of compound and z its charge) and individualizing molecules ionized by the loss or gain of electric charges. Each atom has one or more isotopes of different masses by definition (Chahrour et al., 2015). Thus, the proportion of each isotope observed on a mass spectrum, that is to say the isotopic mass, is characteristic of the presence of certain atoms and their number in the measured ion (Brun et al., 2007). The molecular weight m corresponds to the elemental composition (chemical formula) of the

Table 1: Tables of amino acid.

Symbol	Formula	Monoisotopic	Average
Ala A	Alanine $C_3 H_5 NO$	71.03711	71.0788
Arg R	Arginine $C_6 H_{12} N_4 O$	156.10111	156.1876
Asn N	Asparagine $C_4 H_6 N_2 O_2$	114.04293	114.1039
Asp D	Aspartic Acid $C_4 H_5 N O_3$	115.02694	115.0886
Cys C	Cystine $C_3 H_5 N O_5$	103.00919	103.1448
Gln Q	Glutamine $C_5 H_8 N_2 O_2$	128.05858	123.1308
Glu E	Glutamic Acid $C_5 H_7 N O_3$	129.04259	129.1155
Gly G	Glycine $C_2 H_3 NO$	57.02146	57.0520
His H	Histidine $C_6 H_7 N_3 O$	137.05891	137.1412
Ile I	Isoleucine $C_6 H_{13} NO_2$	113.08406	113.1595
Leu L	Leucine $C_6 H_{13} NO_2$	113.08406	113.1595
Lys K	Lysine $C_6 H_{12} N_2 O$	128.09496	128.1742
Met M	Methionine $C_5 H_9 NOS$	131.04049	131.1986
Phe F	Phenylalanine $C_9 H_9 NO$	147.06841	147.1766
Pro P	Proline $C_5 H_7 NO$	97.05276	97.1167
Ser S	Serine $C_3 H_5 N O_2$	87.03203	87.0782
Thr T	Threonine $C_4 H_7 N O_2$	101.04768	101.1051
TrpW	Tryptophan $C_{11} H_{10} N_2 O$	186.07931	186.2133
Tyr Y	Tyrosine $C_9 H_9 N O_2$	163.06333	163.1760
Val V	Valine $C_5 H_9 NO$	99.06841	99.1326

molecular ion (Kumar Trivedi, 2016). This ion composed of atomic elements is represented by a more or less broad distribution of different isotopes (Brun et al., 2009). Natural isotopes provide an engaging alternative to the accuracy of identification methods (Daron et al., 2016). They have almost the same properties as their parental element, they only differ in the number of neutrons, which explains the difference in mass (Kristjansdottir et al., 2012). Regardless of the similarity in the chemical properties, the presence of the natural isotope promotes independent assessment of molecules because of the mass difference. Here we focused on how considering natural isotopes can offer a highly precise quantification to avoid peptide misidentification.

2 METHODS

Carbon, hydrogen, oxygen, nitrogen and sulfur are the common atoms in amino acids (Gray et al., 1970). They have more than one isotope in nature with different abundance (Table 2).

Carbon and nitrogen are the abundant atoms in peptides with ^{13}C and ^{15}N as the predominant isotopes, followed by sulfur and oxygen isotopes which are present to a lesser extent (Perras et al., 2016). Taking advantage of the isotopic information can provide an absolute quantitation leading to an accurate peptide identification and validation (Hanke et al., 2008;

Table 2: List of most used isotope.

Element	Symbol	Exact mass	Abundance
Hydrogen	1H	1.00783	99.99
	2H	2.01410	0.01
Carbon	^{12}C	12.0000	98.91
	^{13}C	13.0034	1.09
Oxygen	^{16}O	15.9949	99.76
	^{17}O	16.9991	0.04
	^{18}O	17.9992	0.20
Nitrogen	^{14}N	14.0031	99.6
	^{15}N	15.0001	0.4
Sulphur	^{32}S	31.9721	95.02
	^{33}S	32.9715	0.76
	^{34}S	33.9679	4.22
	^{36}S	35.9670	0.02

Costas-Rodriguez et al., 2016). The idea is, instead of representing a peptide by a single mass, peptides need to be represented as a group of different masses balanced to the natural abundance of the natural isotopes.

For an atom, carbon for example, it will be represented by the average mass and not the exact mass. The exact mass of a carbon atom is 12 u, while the average mass is calculated as following: $12 \times 0.9891 + 13 \times 0.0109 = 12.0109$ u. The mass difference for an atom is small, but within a biomolecule (amino acids or peptide), more the number of atoms is large, less this difference is insignificant.

For molecular ions, the situation is more complex since it will take into account the mixed isotopes of each element. The superscript at the upper left is used to indicate the mass number and the subscript at the lower right indicates the atomic number.

For the amino acid phenylalanine $C_9 H_{12} N O_2$:

- The lightest mass corresponds to that of the principal most abundant isotope of each element (monoisotopic mass) (^{12}C , 1H , ^{14}N , ^{16}O):
 $M_{mono} = 9 \cdot 12 + 12 \cdot 1.007825032 + 1 \cdot 14.003074 + 2 \cdot 15.9949146 = 166.0868 \text{ u}$
- For ions of higher mass, we can distinguish those whose weight comes right after the first isotope:
 $^{12}C_8 \ ^{13}C_1 \ ^1H_{12} \ ^{14}N \ ^{16}O_2$, $^{12}C_9 \ ^1H_{11} \ ^2H_1 \ ^{14}N \ ^{16}O_2$, $^{12}C_9 \ ^1H_{12} \ ^{15}N \ ^{16}O_2$, $^{12}C_9 \ ^1H_{12} \ ^{14}N \ ^{16}O_1 \ ^{17}O_1$;
- For the third isotope: $^{12}C_7 \ ^{13}C_2 \ ^1H_{12} \ ^{14}N \ ^{16}O_2$, $^{12}C_9 \ ^1H_{10} \ ^2H_2 \ ^{14}N \ ^{16}O_2$, ,without forgetting the cross products: $^{12}C_8 \ ^{13}C_2 \ ^1H_{11} \ ^2H_1 \ ^{14}N \ ^{16}O_2$

The number of combinations to assemble the isotopes of the same element must be taken into account to go up to the total abundance. For the second isotope: there are nine ways to assemble eight atoms of ^{12}C with one atom of ^{13}C , twelve so as to assemble eleven atoms of 1H and one atom of 2H Thus, we can reconstruct step by step all the isotopes of an ion and construct our collection of different masses for each peptide.

3 RESULTS AND DISCUSSION

The comparison between the average mass and the mass of the first isotope showed a low difference in the case of small molecules, Phenylalanine for example, but this difference increases as the number of elements in the molecules increase.

For a protein molecule, with a mass 23253 u, the gap reaches 14 u (Figure 2).

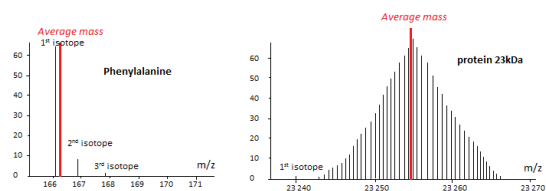


Figure 2: Variance between average mass and monoisotopic mass for an amino acid and a protein.

The first peak of the isotopic profile, often the most abundant for low masses, is the monoisotopic mass

that takes into account just the masses of the most stable isotopes (^{12}C , 1H , ^{16}O , ^{32}S , ^{14}N).

The other peaks all contain at least one heavy isotope of element. In this case of peptide, ^{13}C which is widely represented in biomolecules (of the order of 1 per cent) is mainly responsible for the distribution (range 1 dalton between peaks). It will be observed for the first isotope, a peptide having only ^{12}C ; for the second isotope, the same peptide with ^{13}C , the third with two isotope ^{13}C ... The average mass of a biomolecules takes into account the presence of light isotopes and heavy ones. It corresponds theoretically to the sum of the average weights of each of the elements. The average mass is also the centroid of the masses of the peaks forming the isotopic profile. This mass variance creates bias in results, owing to: misidentification; quantification of an unidentified peptide instead of the expected one; limits the number of quantified peptides; inconsistency between the calculated ratio and the ratio of the peptides composition.

Accurate identification requires consideration of natural isotopes present in peptides. Using the whole isotopic information while calculating the peptide ratio is more precise than using only the monoisotopic peak (Li et al., 2003). The isotopic collection should be computed for each given peptide sequence.

4 CONCLUSION

MS has considerably evolved with the advent of tandem mass spectrometers that allow acquisition of huge quantities of spectra. Thereby, development of meticulous methods for identification and validation of sequence matches will empower this technology. Natural isotopes have turned into an adaptable tool for mass spectrometry studies. This paper has described the properties of natural isotopes and focused on their consideration during peptide quantification. To achieve an optimal absolute quantitation, each peptide to be measured requires an isotopic collection, making it a laborious approach for large proteomic studies. Tools providing exact matches of sequences exist, but we still stand in need for methods providing valid matches for spectra with poor quality. Strong and precise approaches to evaluate the quality of spectra are crucial in order to decrease false positive rates and increase accuracy. Most existing tools are accurate in the case of peptide following a normal fragmentation, but fail for abnormal peptide. Fragmentation is highly dependent on physics and chemistry of amino acids and peptide sequences. Natural isotopes exploitation is also valuable for peptide iden-

tification and validation by facilitating the identification of peptide containing mutations, posttranslational modifications and/or abnormal fragmentation. Our study showed that calculating peptides masses using the whole isotopic collection is more precise than using only the monoisotopic masses. Hereby, considering natural isotope would appear to satisfy the criteria for an optimal quantitative mass spectrometry strategy leading to an accurate peptide/protein identification.

REFERENCES

- Aebersold, R. and Mann, M. (2003). Mass spectrometry-based proteomics. *Nature*, 422(6928):198–207.
- Brun, V., Dupuis, A., Adrait, A., Marcellin, M., Thomas, D., Court, M., Vandenesch, F., and Garin, J. (2007). Isotope-labeled Protein Standards: Toward Absolute Quantitative Proteomics. *Molecular & Cellular Proteomics*, 6(12):2139–2149.
- Brun, V., Masselon, C., Garin, J., and Dupuis, A. (2009). Isotope dilution strategies for absolute quantitative proteomics. *Journal of Proteomics*, 72(5):740–749.
- Chahrour, O., Cobice, D., and Malone, J. (2015). Stable isotope labelling methods in mass spectrometry-based quantitative proteomics. *Journal of Pharmaceutical and Biomedical Analysis*, 113:2–20.
- Costas-Rodriguez, M., Delanghe, J., and Vanhaecke, F. (2016). High-precision isotopic analysis of essential mineral elements in biomedicine: natural isotope ratio variations as potential diagnostic and/or prognostic markers. *TrAC Trends in Analytical Chemistry*, 76:182–193.
- Daron, M., Blamart, D., Peral, M., and Affek, H. (2016). Absolute isotopic abundance ratios and the accuracy of 47 measurements. *Chemical Geology*, 442:83–96.
- Deutsch, E. W., Mendoza, L., Shteynberg, D., Farrah, T., Lam, H., Tasman, N., Sun, Z., Nilsson, E., Pratt, B., Prazen, B., Eng, J. K., Martin, D. B., Nesvizhskii, A. I., and Aebersold, R. (2010). A guided tour of the Trans-Proteomic Pipeline. *PROTEOMICS*, 10(6):1150–1159.
- Gerber, S. A., Rush, J., Stemman, O., Kirschner, M. W., and Gygi, S. P. (2003). Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proceedings of the National Academy of Sciences*, 100(12):6940–6945.
- Gray, W. R., Wojcik, L. H., and Futrell, J. H. (1970). Application of mass spectrometry to protein chemistry. II. Chemical ionization studies on acetylated permethylated peptides. *Biochemical and Biophysical Research Communications*, 41(5):1111–1119.
- Hanke, S., Besir, H., Oesterhelt, D., and Mann, M. (2008). Absolute SILAC for Accurate Quantitation of Proteins in Complex Mixtures Down to the Attomole Level. *Journal of Proteome Research*, 7(3):1118–1130.
- Hunt, D. F., Yates, J. R., Shabanowitz, J., Winston, S., and Hauer, C. R. (1986). Protein sequencing by tandem mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America*, 83(17):6233–6237.
- Koehler, C. J., Arntzen, M., Strozynski, M., Treumann, A., and Thiede, B. (2011). Isobaric Peptide Termini Labeling Utilizing Site-Specific N-Terminal Succinylation. *Analytical Chemistry*, 83(12):4775–4781.
- Kristjansdottir, K., Takahashi, S., L., S., and J., S. (2012). Strategies and Challenges in Measuring Protein Abundance Using Stable Isotope Labeling and Tandem Mass Spectrometry. In Prasain, J., editor, *Tandem Mass Spectrometry - Applications and Principles*. Intech.
- Kumar Trivedi, M. (2016). Gas Chromatography-Mass Spectrometric Analysis of Isotopic Abundance of ^{13}C , ^2H , and ^{18}O in Biofield Energy Treated p-tertiary Butylphenol (PTBP). *American Journal of Chemical Engineering*, 4(4):78.
- Li, X.-j., Zhang, H., Ranish, J. A., and Aebersold, R. (2003). Automated Statistical Analysis of Protein Abundance Ratios from Data Generated by Stable-Isotope Dilution and Tandem Mass Spectrometry. *Analytical Chemistry*, 75(23):6648–6657.
- Listgarten, J. and Emili, A. (2005). Statistical and Computational Methods for Comparative Proteomic Profiling Using Liquid Chromatography-Tandem Mass Spectrometry. *Molecular & Cellular Proteomics*, 4(4):419–434.
- Mann, M. and Kelleher, N. L. (2008). Precision proteomics: The case for high resolution and high mass accuracy. *Proceedings of the National Academy of Sciences*, 105(47):18132–18138.
- Perras, F. A., Chaudhary, U., Slowing, I. I., and Pruski, M. (2016). Probing Surface Hydrogen Bonding and Dynamics by Natural Abundance, Multidimensional, ^{17}O DNP-NMR Spectroscopy. *The Journal of Physical Chemistry C*, 120(21):11535–11544.
- Pratt, J. M., Petty, J., Riba-Garcia, I., Robertson, D. H. L., Gaskell, S. J., Oliver, S. G., and Beynon, R. J. (2002). Dynamics of protein turnover, a missing dimension in proteomics. *Molecular & cellular proteomics: MCP*, 1(8):579–591.
- Schmidt, A., Forne, I., and Imhof, A. (2014). Bioinformatic analysis of proteomics data. *BMC Systems Biology*, 8(Suppl 2):S3.
- Yates, J. R., Ruse, C. I., and Nakorchevsky, A. (2009). Proteomics by Mass Spectrometry: Approaches, Advances, and Applications. *Annual Review of Biomedical Engineering*, 11(1):49–79.