

Mass-Based Classification (MBC) of Peptides: Highly Accurate Precursor Ion Mass Values Can Be Used to Directly Recognize Peptide Phosphorylation

Bernhard Spengler and Alfons Hester

Institute of Inorganic and Analytical Chemistry, Justus Liebig University Giessen, Giessen, Germany

Accurate mass values as obtainable by Fourier transform ion cyclotron resonance mass spectrometry (FTICR-MS) were employed in a theoretical study to differentiate between nonmodified and phosphorylated peptides. It was found that for peptide masses up to 1000 u more than 98% of all theoretical monophosphorylated peptides (all possible combinations of proteinogenic amino acids having one phosphorylation on S, T, or Y) can be distinguished from nonphosphorylated peptides directly by their mass, if mass values are determined with an accuracy of better than ± 0.1 ppm. At a peptide mass of 1500 u still 70% of all possible monophosphorylated peptides are distinguishable from nonmodified peptides by their accurate mass alone. In contrast to established techniques of data-dependent multidimensional mass spectrometry, only the mass of the precursor ion is necessary to decide upon subsequent fragment ion analysis of a peptide for sequence analysis in an LC-MS/MS investigation of a complex sample, when using a precalculated mass distribution table of theoretical peptides. A mass distribution table of nonphosphorylated and monophosphorylated peptides with a bin width of 0.1 mu was made available via the open web site www.peptidecomposer.com. (J Am Soc Mass Spectrom 2008, 19, 1808–1812) © 2008 Published by Elsevier Inc. on behalf of American Society for Mass Spectrometry

Accurate masses of molecules or molecular ions build up a discrete structure in mass space, as a result of the mass defect of the isotopes they are composed of. Masses of atoms are therefore not the sum of masses of a certain number of nuclear protons and neutrons plus the sum of masses of electrons, but include an additional mass deficit due to the specific nuclear binding energies of the atom. As a result, each isotopic species carries a characteristic mass signature expressed as an isotope-specific accurate mass value and furthermore, each molecule carries digital signatures (i.e., number and accurate mass) of their atomic building blocks. Regarding the mass space of different substance classes such as peptides, carbohydrates, or lipids, varying ratios of their common elements (CHNOS) lead to accurate-mass distributions, which do not necessarily overlap and therefore can eventually be used for substance class identification (Figure 1a). If elemental compositions of different substance class molecules are identical accidentally, however, accurate masses are identical and cannot be resolved even with the highest mass resolving power.

When introducing a new element, on the other hand,

accurate mass distributions definitely become distinguishable, since the mass defect of the added element results in a new signature of the accurate masses of the molecules. This is the case, for example, when comparing nonmodified and phosphorylated peptides. With a sufficiently high mass resolving power, such species can always be distinguished from each other.

As instruments for determining molecular masses are getting more and more accurate with higher and higher mass resolution, investigation of this structure is becoming an important and fundamental scientific goal. Today, mass accuracies in the range of ± 0.2 ppm can routinely be achieved with internal calibration on ion trap Fourier transform ion cyclotron resonance mass spectrometers [1–4], and it is quite likely that the ± 0.1 ppm range of mass accuracies will be accessible in the foreseeable future even with external calibration.

The structure of molecular masses can be distinguished into a coarse, fine, and hyperfine structure, similar to the spectroscopic structure of molecules. Introducing a phosphorylation, for example, leads to a shift of the distribution to a higher mass deficit (to smaller masses), but does not create a separation of the distribution of phosphorylated peptides into the “forbidden zone” of the distribution of unmodified peptides (Figure 1b).

Address reprint requests to Dr. B. Spengler, Institute of Inorganic and Analytical Chemistry, Justus Liebig University Giessen, Schubertstr. 60, D-35392 Giessen, Germany. E-mail: bernhard.spengler@anorg.chemie.uni-giessen.de; <http://www.uni-giessen.de/analytik/>

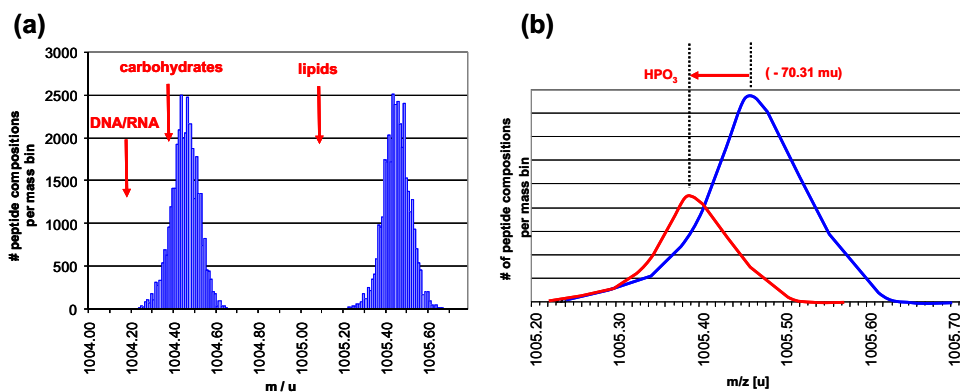


Figure 1. (a) Mass distribution of theoretical peptides for nominal mass 1004 u and 1005 u (blue) and center masses of distributions of oligonucleotides, carbohydrates and lipids. Lipidated peptides, for example, fall into the area between lipids (higher accurate mass) and unmodified peptides (lower accurate mass). (b) Shift of mass distribution to lower accurate masses by monophosphorylation. Peptides with one phosphorylation still overlap with the mass distribution of unmodified peptides in low-resolution mode, but can eventually be distinguished by high-resolution mass evaluation.

In high-resolution evaluation, however, phosphorylated and nonphosphorylated peptides do not overlap any more. It is then just a question of instrumental mass resolving power and accuracy, whether neighboring mass signals can be unequivocally attributed to phosphorylation or nonphosphorylation (Figure 2).

Knowledge of the structure of molecular masses can be used in various ways, including autocalibration of mass spectrometric instruments [2], mass-based classification of biomolecular samples, composition-based de novo sequencing of peptides [2–4], or structure analysis of organic molecules. In the following study, the discrimination capacity of accurate mass values for recognizing peptide phosphorylations is described, and results are made available for analytical use via internet access.

Experimental

Calculation of Peptide Ion Mass Values

Isotope mass values of ^1H , ^{12}C , ^{14}N , ^{16}O , ^{31}P , and ^{32}S used for calculation follow the 1995 table of Audi and Wapstra [5]:

$$\begin{aligned} m(^1\text{H}) &= 1.00782503214 \text{ u} \\ m(^{12}\text{C}) &= 12.00000000000 \text{ u} \\ m(^{14}\text{N}) &= 14.00307400524 \text{ u} \\ m(^{16}\text{O}) &= 15.9949146221 \text{ u} \\ m(^{31}\text{P}) &= 30.973761512 \text{ u} \\ m(^{32}\text{S}) &= 31.972070690 \text{ u} \\ \text{Electron mass value used was} \\ m_e &= 0.0005485799 \text{ u} \end{aligned}$$

Revised values of isotopic masses [6] were found to lead to identical results in this study.

Monoisotopic mass values of singly charged peptides $[\text{M} + \text{H}]^+$ were calculated as the sum of amino acid mass increments of Table 1 and the mass of H_3O^+ :

$$\begin{aligned} M(\text{H}_3\text{O}^+) \\ = 3 * m(^1\text{H}) + m(^{16}\text{O}) - m_e = 19.0178411386 \text{ u}. \end{aligned}$$

Mass values of all possible amino acid combinations of Table 1 were calculated by exhaustive enumeration of all possible amino acid sum formulae. A counting list of abundances of peptide compositions between 50 and 2000 u with a bin width of 0.0001 u (= 0.1 mu) was filled with mass hits within ± 0.05 mu of the bin value. Leucine and isoleucine were handled as equal and were not counted separately. Monophosphorylated peptides were calculated for peptide compositions, including one phosphorylation on either a serine, threonine, or tyrosine amino acid.

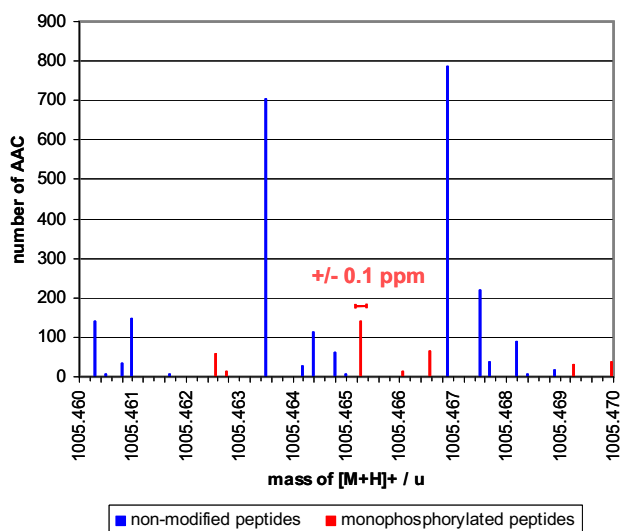


Figure 2. Phosphorylated peptides have accurate masses distinguishable from those of unmodified peptides. A sufficiently high mass resolving power allows identifying phosphorylation directly from the accurate precursor ion mass. AAC is amino acid combinations (i.e., peptide compositions).

Table 1. Accurate values of monoisotopic amino acid mass increments as used in this study. Calculations are based on isotopic mass values of ^1H , ^{12}C , ^{14}N , ^{16}O , ^{31}P , and ^{32}S [5]

Amino acid	code		Avg. mass/ Elemental composition	Monoisotopic mass increment/u
Alanine	A	Ala	$\text{C}_3\text{H}_5\text{NO}$	71.03711378804
Cysteine	C	Cys	$\text{C}_3\text{H}_5\text{NOS}$	103.00918447804
Aspartic acid	D	Asp	$\text{C}_4\text{H}_5\text{NO}_3$	115.02694303224
Glutamic acid	E	Glu	$\text{C}_5\text{H}_7\text{NO}_3$	129.04259309652
Phenylalanine	F	Phe	$\text{C}_9\text{H}_9\text{NO}$	147.0684139166
Glycine	G	Gly	$\text{C}_2\text{H}_3\text{NO}$	57.02146372376
Histidine	H	His	$\text{C}_6\text{H}_7\text{N}_3\text{O}$	137.0589118628
Isoleucine	I	Ile	$\text{C}_6\text{H}_{11}\text{NO}$	113.08406398088
Lysine	K	Lys	$\text{C}_6\text{H}_{12}\text{N}_2\text{O}$	128.09496301826
Leucine	L	Leu	$\text{C}_6\text{H}_{11}\text{NO}$	113.08406398088
Methionine	M	Met	$\text{C}_5\text{H}_9\text{NOS}$	131.0404846066
Asparagine	N	Asn	$\text{C}_4\text{H}_6\text{N}_2\text{O}_2$	114.04292744752
Proline	P	Pro	$\text{C}_5\text{H}_7\text{NO}$	97.05276385232
Glutamine	Q	Gln	$\text{C}_5\text{H}_8\text{N}_2\text{O}_2$	128.0585775118
Arginine	R	Arg	$\text{C}_6\text{H}_{12}\text{N}_4\text{O}$	156.10111102874
Serine	S	Ser	$\text{C}_3\text{H}_5\text{NO}_2$	87.03202841014
Threonine	T	Thr	$\text{C}_4\text{H}_7\text{NO}_2$	101.04767847442
Valine	V	Val	$\text{C}_5\text{H}_9\text{NO}$	99.0684139166
Tryptophan	W	Trp	$\text{C}_{11}\text{H}_{10}\text{N}_2\text{O}$	186.07931295398
Tyrosine	Y	Tyr	$\text{C}_9\text{H}_9\text{NO}_2$	163.0633285387
p-Serine	pS	pSer	$\text{C}_3\text{H}_6\text{NO}_5\text{P}$	166.99835882058
p-Threonine	pT	pThr	$\text{C}_4\text{H}_8\text{NO}_5\text{P}$	181.01400888486
p-Tyrosine	pY	pTyr	$\text{C}_9\text{H}_{10}\text{NO}_5\text{P}$	243.02965894914

It has to be noted that the resulting abundance list describes the numbers of possible peptide compositions (i.e., amino acid sum formulae). The number of possible

peptide sequences within a certain mass bin is of course much higher.

All calculations were performed with double precision floating point arithmetic (64 bit), using the Peptide Composer software [2, 3, 4, 7].

Results

Nonmodified Peptides

A total of 8.9 billion nonmodified peptide compositions was calculated for peptide masses between 50 and 2000 u. The calculated accurate mass values of monoisotopic singly charged protonated peptides were sorted into 0.1 mu wide bins between mass 50 and 2000 u. Bins became populated between zero and a maximum of 202,732 compositions.

Differentiation between nonmodified peptides and modified peptides or other substance classes first requires the existence of empty bins within the mass distribution of the nonmodified peptides. The portion of empty (nonpopulated) bins as a function of mass was determined and is displayed in Figure 3.

Monophosphorylated Peptides

A total of 4.6 billion monophosphorylated peptide compositions was calculated for peptide masses between 50 and 2000 u (supplementary material S1, which can be found in the electronic version of this article). Their number is rising almost exponentially with mass.

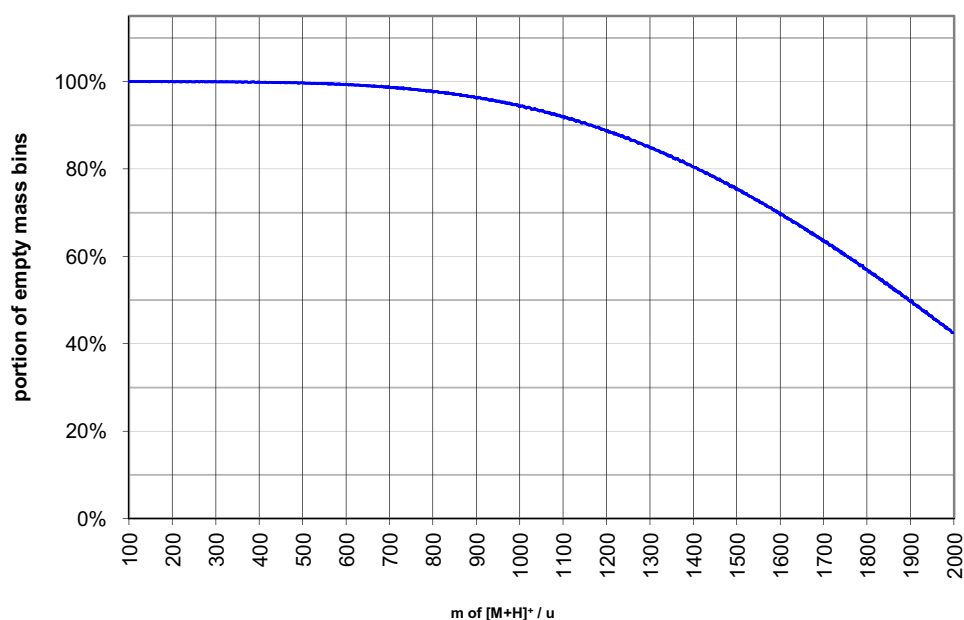


Figure 3. Portion of nonpopulated 0.1 mu wide mass bins of singly charged protonated peptides as a function of peptide ion mass. The distribution shows that for example at mass 1000 u, about 95% of the mass bins are unpopulated leaving mass space for modified peptides or other substance classes to be directly distinguished from nonmodified peptides by the accurate precursor ion mass alone.

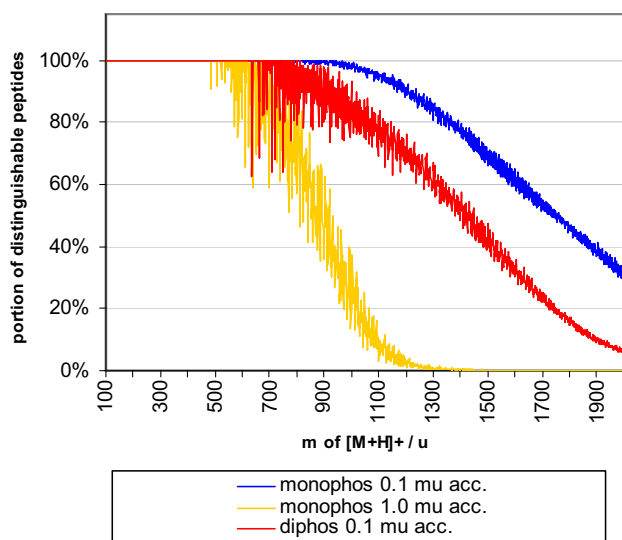


Figure 4. Portion of mono- and diphosphorylated peptides that can be distinguished from nonmodified peptides by accurate precursor ion mass if mass accuracy is better than ± 0.1 mu (± 0.1 ppm at mass 1000 u; blue and red curve) or better than ± 1.0 mu (± 1.0 ppm at mass 1000 u; yellow curve). Monophosphorylated peptides can be distinguished easier from nonmodified peptides than diphosphorylated peptides.

Distinguish Monophosphorylated Peptides

The portion of mono- and diphosphorylated peptides that can be distinguished by accurate mass from nonmodified peptides was counted for an assumed instrumental accuracy of 0.1 mu (0.1 ppm at mass 1000 u) and 1.0 mu (1 ppm at mass 1000), respectively. The results are shown in Figure 4. As can be seen, the discrimination between modified and unmodified peptides is considerably better for the monophosphorylated peptides compared with diphosphorylated peptides. Furthermore, it is obvious that the highest available mass accuracy is necessary to make use of mass-based discrimination effects.

Access to Database

The mass distribution database of unmodified and monophosphorylated peptides is available online through the web site <http://www.peptidecomposer.com> (page "Web access"). The form requires input of the accurate mass and the width of the mass range to be displayed:

monoisotopic mass of $[M + H]^+$
 valid masses: 50u to 2000u
 mass range \pm
 Example 1234.5678

Output Settings Graphics in black/white
 Show accuracy bar in graphics

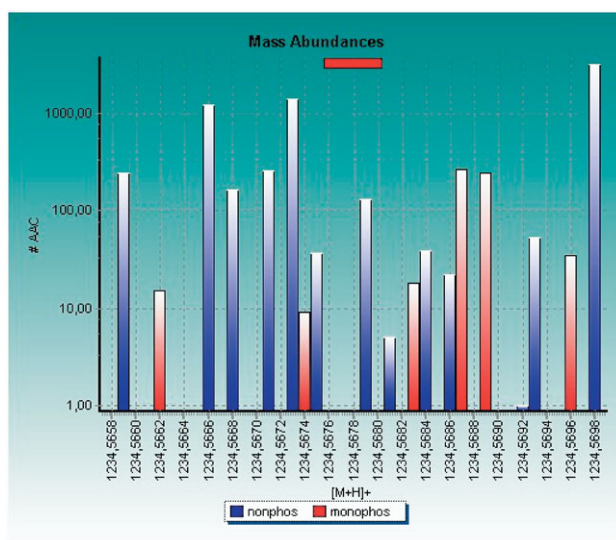


Figure 5. Logarithmic representation of a selected mass distribution with a mass range of ± 0.002 u. The red horizontal bar represents the chosen instrumental mass accuracy of ± 0.2 ppm for comparison.

Output is a graphical representation of the distribution in either linear (supplementary material S2) or logarithmic scaling (Figure 5), and a listing of the bin populations for unmodified and monophosphorylated peptides. An instrumental mass accuracy can be specified, to display an accuracy bar in the diagram for comparison.

The coarse distribution showing the well-known "islands" and "forbidden zones" [8] of possible peptide masses can be visualized when setting the mass range to, e.g., ± 3 u (supplementary material S3).

Conclusions

Calculation of accurate masses of all possible peptide compositions by combinatorial combination of the 20 proteinogenic amino acids led to a distribution table for nonmodified and monophosphorylated peptides. Access to the table is available through the internet. The results indicate that distinguishing monophosphorylated and nonphosphorylated peptides based on the accurate precursor ion mass alone is possible if the instrumental mass accuracy is sufficiently high. For smaller peptides up to about 700 u, an instrumental

u \pm ppm

u
 The . (dot) is the decimal separator.
 Please do not use thousands separator.

mass accuracy of ± 1 ppm sufficient to identify phosphorylation and peptide composition. For peptides above 1300 u, a minimum mass accuracy of 0.1 ppm is necessary to efficiently discriminate peptide compositions and phosphorylation states.

Classification of unknown compounds by precursor mass evaluation is possible using the mass distribution table, when deciding whether a certain mass signal belongs to a peptide, an oligonucleotide, or a lipid, for example.

Instrumental capabilities of modern mass spectrometers are now high enough to access the discrete structure of accurate molecular masses. A new type of bioinformatics tools therefore is evolving, having a strong potential for fast and highly informative bioanalytical evaluation.

Acknowledgments

The authors gratefully acknowledge financial support by the European Union (STREP project LSHG-CT-2005-518194), by the

Bundesministerium für Bildung und Forschung (NGFN-Project 0313442), and by the Deutsche Forschungsgemeinschaft (SP 314-10-1).

References

1. Smith, R. D.; Anderson, G. A.; Lipton, M.S.; Pasa-Tolic, L.; Shen, Y.; Conrads, Th. P.; Veenstra, T.D.; Udseth, H.R. An Accurate Mass Tag Strategy for Quantitative and High-Throughput Proteome Measurements. *Proteomics* **2002**, *2*, 513–523.
2. Spengler, B. De Novo Sequencing, Peptide Composition Analysis, and Composition-Based Sequencing: A New Strategy Employing Accurate Mass Determination by Fourier Transform Ion Cyclotron Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **2004**, *15*, 703–714.
3. Thieu, V. A.; Kirsch, D.; Flad, T.; Müller, C.; Spengler, B. Direct Protein Identification from Nonspecific Peptide Pools by High-Accuracy MS Data Filtering. *Angew. Chem. Int. Ed.* **2006**, *45*, 3317–3319.
4. Spengler, B. Accurate Mass as a Bioinformatics Parameter in Data-to-Knowledge Conversion: Fourier Transform Ion Cyclotron Resonance Mass Spectrometry for peptide de novo sequencing. *Eur. Mass Spectrom.* **2007**, *13*, 83–87.
5. Audi, G.; Wapstra, A. H. *Nucl. Phys.* **1995**, *A595* 409–480.
6. Audi, G.; Wapstra, A. H.; Thibault, C. *Nucl. Phys.* **2003**, *A729*, 337–676.
7. Spengler, B. Peptide Composer, Calculation Software Package for Combinatorial Calculation of Peptide Mass Values, 2008 (www.peptidecomposer.com).
8. Mann, M. Useful Tables of Possible and Probable Peptide Masses. *Proceedings of the 43rd ASMS Conference on Mass Spectrometry and Allied Topics*; Atlanta, GA, May, 1995; p 639.