Received: 17 June 2008 ■ Revised: 11 September 2008 ■ Accepted: 12 September 2008 ■ Publication: 22 September 2008

EJMS EUROPEAN JOURNAL OF MASS SPECTROMETRY

OPEN ACCESS

Protease-dependent fractional mass and peptide properties

Harald Barsnes,[a,*] Ingvar Eidhammer,[a] Véronique Cruciani[b,c] and Svein-Ole Mikalsen[b]

[a]Department of Informatics, University of Bergen, Norway, E-mail: Harald.Barsnes@ii.uib.no

[b]Institute for Cancer Research, Rikshospitalet-Radiumhospitalet Medical Center, Oslo, Norway

[c]Present address: Affitech AS, Oslo Research Park, Oslo, Norway

Mass spectrometric analyses of peptides mainly rely on cleavage of proteins with proteases that have a defined specificity. The specificities of the proteases imply that there is not a random distribution of amino acids in the peptides. The physico–chemical effects of this distribution have been partly analyzed for tryptic peptides, but to a lesser degree for other proteases. Using all human proteins in Swiss-Prot, the relationships between peptide fractional mass, pI and hydrophobicity were investigated. The distribution of the fractional masses and the average regression lines for the fractional masses were similar, but not identical, for the peptides generated by the proteases trypsin, chymotrypsin and gluC, with the steepest regression line for gluC. The fractional mass regression lines for individual proteins showed up to ±100 ppm in mass difference from the average regression line and the peptides generated showed protease-dependent properties. We here show that the fractional mass and some other properties of the peptides are dependent on the protease used for generating the peptides. With the increasing accuracy of mass spectrometry instruments, it is possible to exploit the information embedded in the fractional mass of unknown peaks in peptide mass fingerprint spectra.

Keywords: peptides, MALDI, accurate mass, fractional mass, peptide mass fingerprint

Introduction

Protein identification and identification of post-translational modifications are major tasks in proteomics, mainly by employing mass spectrometry (MS). Commonly, the unknown protein is digested with a protease and the resulting peptides are analyzed in an MS instrument. In matrix-assisted laser desorption/ionization time-of-flight (MALDI-ToF) instruments, the *m/z* values of (mainly) singly charged and intact peptides are measured, resulting in a peptide mass fingerprint (PMF). The PMF can be used for searching databases and obtaining lists of candidate proteins.[1–3] Trypsin is by far the most common protease, but other proteases are also used.

The masses of the atoms occurring in amino acids and their post-translational modifications are all close to an integer (monoisotopic masses: H, 1.00783; C, 12.0000; N, 14.0031; O, 15.9949; P, 30.9738; S, 31.9721). Therefore, the masses of the amino acids are also close to an integer. As noted by Mann,[4] this implies that peptide masses distribute in clusters with a mass difference of slightly more than 1 Da. Thus, if the peptide masses are plotted against the fractional masses (i.e. the numbers after the decimal sign), a slope of approximately 0.000455–0.000495 is obtained.[4–6] By definition, the fractional mass is between 0 and 1. Its distribution is discontinuous at a mass of approximately 2000 where the sum of the fractional masses of the individual amino acids in the peptide exceeds 1. To avoid the discontinuity, the term deltamass, the sum of the fractional masses, has been used.[7] We will use fractional mass as a generic term and deltamass as a specific term when the sum of the amino acids' fractional masses can exceed 1.

The mass clustering effect has some practical implications. It makes it possible to remove peaks that are unlikely to be

ISSN: 1469-0667
doi: 10.1255/ejms.934

peptides, for example, matrix–alkali ions and thereby improve protein identification,[8–10] or these ions can be distinguished and used in calibration.[11] The average fractional mass can be used in a pre-calibration of MS spectra, significantly reducing mass errors,[6] or as an extra control of proper calibration.[12] The fractional mass may help to predict whether an unknown peak is due to glycosylation.[7] It is theoretically possible to calculate the elemental composition of a peptide from its mass, but due to the mass clustering effect this can only be done for small peptides using the most accurate instruments available.[13,14] However, accurate mass determination may significantly improve MS/MS *de novo* sequencing.[14–16]

Most of the studies mentioned above have used trypsin for the proteolytic cleavage. Tryptic peptides contain one (or more) of the amino acids with the highest fractional masses, arginine and lysine. It has been noted that there is a correlation between high-performance liquid chromatography (HPLC) retention in reverse phase chromatography and peptide fractional mass when a narrow mass range (1 Da) is considered.[7] We have therefore asked how properties often used to separate peptides are correlated to the fractional mass of the peptide. We have further investigated whether the use of other proteases with different cleavage specificities would change these correlations. For this purpose, we have compared the theoretical peptides created by digesting all human proteins in the Swiss-Prot[17] database by trypsin, chymotrypsin and gluC.

Proteins and protelytic digestion

All human proteins were extracted from the Swiss-Prot database (Release 49) using Swissknife,[18] resulting in 13,358 sequences. Sequences fulfilling the two following criteria were excluded: (i) the presence of non-standard amino acids, and (ii) the lack of peptides in the *m/z* range 490–3510. The protein sequences were digested *in silico* by trypsin, chymotrypsin and gluC, using the digestion tool of MassSorter.[19] The following cleavage rules were used: trypsin cleaved C-terminal to arginine and lysine, but not if followed by proline; chymotrypsin cleaved C-terminal to phenylalanine, tyrosine, tryptophan and leucine, but not if followed by proline; and gluC cleaved C-terminal to aspartic acid and glutamic acid, but not if followed by proline. Only unmodified peptides with no missed cleavages were considered. Cysteine was treated as unmodified. The final number of proteins and peptides were 13,260 and 455,835 for trypsin, 13,249 and 588,360 for chymotrypsin and 13,228 and 463,030 for gluC.

For the calculations of the individual fractional mass regression lines of proteins, only proteins yielding 15 or more peptides were used. For this analysis, 9624, 11,219 and 9378 proteins were cleaved by trypsin, chymotrypsin and gluC, respectively. Theoretical pI of the peptides were calculated by pI_tool.[20] The hydrophobicities were based on the GRAVY scale and were calculated as the average of all amino acids in the peptide.[21] Similar results and conclusions were obtained when other scales of hydrophobicity were used (results not shown).

Results

The distributions of fractional masses of proteolytic peptides are shown in Figures 1(a)–(c). At lower peptide masses, the distribution of fractional masses is relatively narrow, but it is wider at higher masses. The average regression line for the tryptic fractional mass (here considered as deltamass to avoid the discontinuity) was

$$DM_{tryp} = 0.0036937 + 0.0004886M_i$$

where DM_{tryp} is the deltamass obtained for the integer mass M_i. The factor 0.0004886 is close to previous results.[4–6] The average regression lines for deltamasses generated by the three proteases trypsin, chymotrypsin and gluC are plotted in Figure 1(d). The chymotryptic regression line was close to that of trypsin,

$$DM_{chymo} = -0.0003614 + 0.0004900M_i.$$

On the other hand, gluC had a steeper regression line,

$$DM_{gluC} = -0.0069709 + 0.0005108M_i.$$

Thus, different enzymes have different regression lines. The regression lines intercept the ordinate at positive values when the C-terminal amino acid has high fractional mass, and at negative values when the C-terminal amino acid has low fractional mass, similar to the recent results obtained by Wolski *et al*.[22]

The regression line for the fractional mass may be used to improve calibration of MS data.[6] We therefore investigated how the regression lines for the individual proteins distributed. A relatively wide distribution was obtained for the three proteases (Figure 2). For example, for tryptic peptides at masses of 1500 and 2000, there were 162 ppm and 185 ppm, respectively, between the lower and the upper regression lines, indicating that, in extreme cases, the use of the average fractional mass regression line may be a disadvantage. However, most proteins showed a relatively narrow distribution around the average regression line and 96.7%, 94.2% and 96.5% of the individual regression lines for trypsin, chymotrypsin and gluC, respectively, were within ±0.05 Da of the average regression line at an integer mass of 1500 (corresponding to ±33.3 ppm).

It has previously been shown that tryptic peptides in the mass range 1968.70–1969.25 (deltamass 0.7–1.25) showed a good correlation with an HPLC retention index with the lower deltamass having lower retention and the higher deltamass having higher retention.[7] The relationships between fractional mass, hydrophobicity and pI were therefore also studied for chymotrypsin and gluC. As expected,[7] tryptic peptides showed a positive correlation between hydrophobicity and increasing fractional mass for all mass ranges studied (499–501, 999–1001, 1499–1501, 1999–2001, 2499–2501, 2999–3001, 3499–3501),

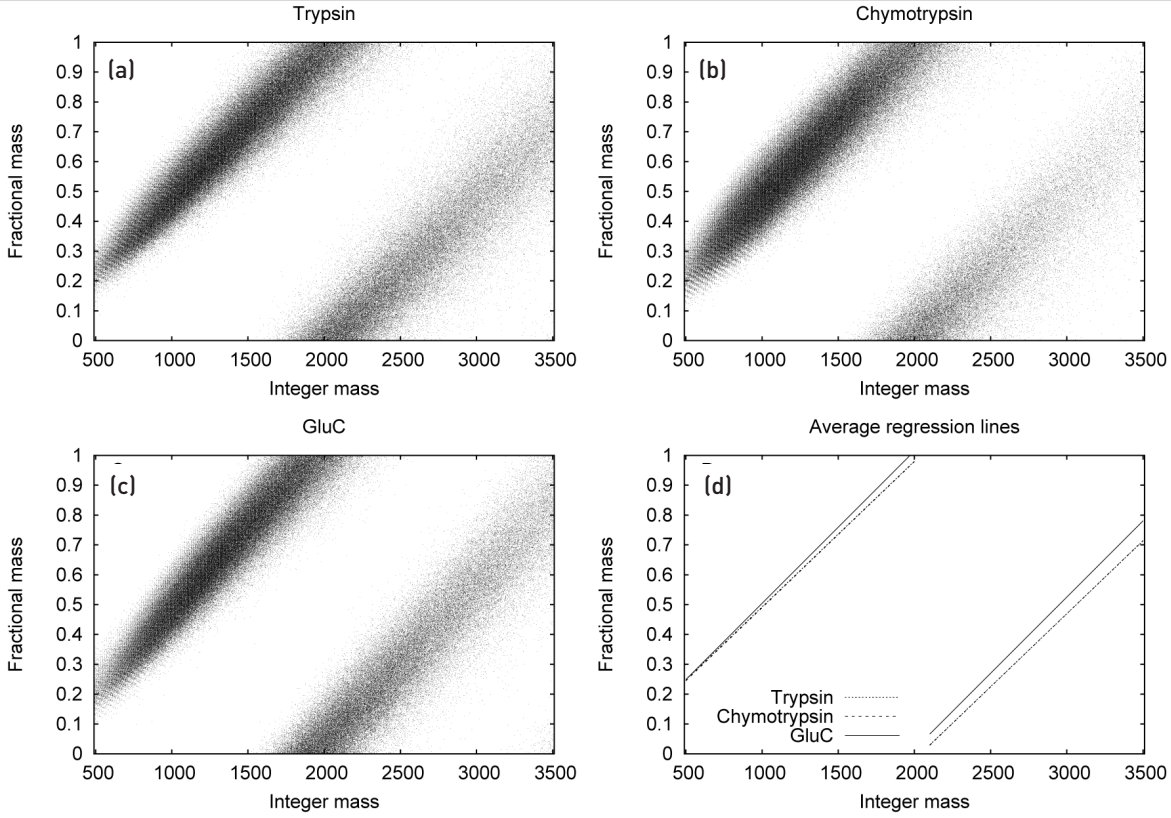

Figure 1. Distribution of fractional masses of proteolytic peptides. Human proteins were extracted from SwissProt and theoretically digested with (a) trypsin, (b) chymotrypsin or (c) gluC, assuming no missed cleavages. The fractional mass was plotted against the integer mass for each of the peptides and the average regression lines were calculated. (d) Plot of the average regression lines. Note that peptides generated by gluC give a steeper regression line than peptides generated by the two other proteases. Note also that trypsin and chymotrypsin have nearly identical regression lines.

with correlation coefficient $r = 0.63$ to 0.78. This is exemplified by the mass range 1499–1501 [Figure 3(a)]. On the other hand, peptides generated by chymotrypsin or gluC [Figures 3(b) and (c)] showed weak correlation between hydrophobicity and fractional mass ($r = -0.38$ to 0.38 for chymotrypsin, except for the mass range 499–501 where $r = 0.61$; gluC had $r = 0.02$ to 0.28), and the slopes of the regression lines in several cases were not significantly different from zero. In other words, considering peptides within a narrow mass range, tryptic peptides with low fractional mass tended to be more hydrophilic and peptides with high fractional mass tended to be more hydrophobic. In contrast, peptides generated by chymotrypsin or gluC had relatively similar hydrophobicities at low and high fractional masses, still within a narrow mass range.

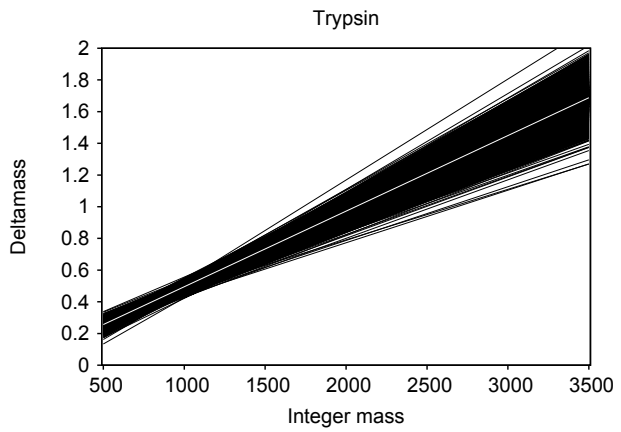

Figure 2. Distribution of individual regression lines for integer mass vs deltamass. The proteins were cleaved in silico by trypsin. Individual regression lines in black with the average regression line in white. Similar results were also found for chymotrypsin and gluC (not shown).

The positioning of certain amino acids within the peptide may affect pI,[23] but most peptides have their experimental pI close to the theoretical pI.[23] As observed before for tryptic peptides,[23,24] the pI tended to distribute into clusters (Figure 3). The pI clustering effect was somewhat more pronounced for trypsin and gluC than for chymotrypsin [Figures 3(a)–(c)]. The strongest positive correlation between fractional mass and pI was found for chymotrypsin ($r = 0.59$ to 0.68) and gluC ($r = 0.36$ to 0.59). Although the regression lines for fractional mass and pI of tryptic peptides in general had weaker correlation coefficients ($r = 0.29$ to 0.51), the slopes

were still significantly different from zero ($p < 0.0001$, except for the mass range 3499–3501). The stronger correlation between pI and fractional mass for chymotrypsin and gluC than for trypsin is a direct consequence of the lower correlation between hydrophobicity and fractional mass as will be explained in the Discussion.

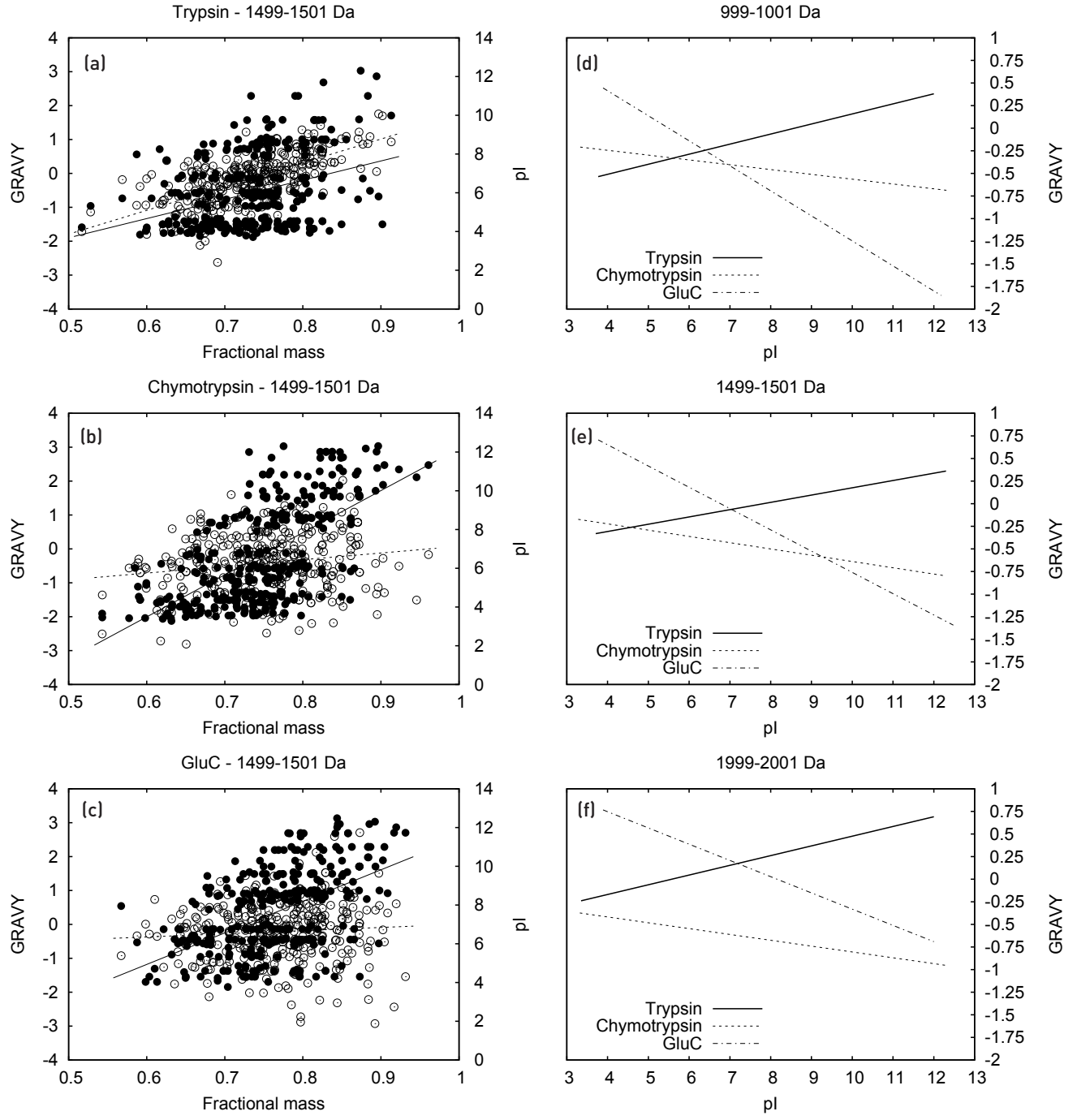

Figure 3. Properties of peptides generated by (a) trypsin, (b) chymotrypsin and (c) gluC. The relationships between fractional mass vs. GRAVY (open circles and dotted line) and fractional mass vs pI (filled circles and unbroken line) in the mass range 1499–1501 for peptides generated by (a) trypsin (319 peptides; $r = 0.64$ for GRAVY and 0.34 for pI; both slopes are significantly different from zero, $p < 0.0001$), (b) chymotrypsin (317 peptides; $r = 0.17$ for GRAVY and 0.64 for pI; both slopes are significantly different from zero, $p \leq 0.002$) and (c) gluC (334 peptides; $r = 0.07$ for GRAVY and 0.51 for pI; the slope for pI is significantly different from zero, $p < 0.0001$). [(d), (e), (f)] Regression lines for the relationship of hydrophobicity vs pI for the three proteases in the mass ranges (d) 999–1001 (trypsin, 541 peptides, $r = 0.24$; chymotrypsin, 1024 peptides, $r = 0.14$; gluC, 608 peptides, $r = -0.59$), (e) 1499–1501 (trypsin, $r = 0.21$; chymotrypsin, $r = 0.20$; gluC, $r = -0.60$) and (f) 1999–2001 (trypsin, 195 peptides, $r = 0.24$; chymotrypsin, 145 peptides, $r = 0.20$; gluC, 194 peptides, $r = -0.52$).

When hydrophobicity was plotted against pI, there was a relatively low, but positive, correlation ($r = 0.17$ to 0.51; with slopes significantly different from zero, $p < 0.001$) for tryptic peptides in all mass ranges, exemplified by 999–1001, 1499–1501 and 1999–2001 [Figures 3(d)–(f)]. Thus, tryptic peptides with low hydrophobicity tended to have low pI and tryptic peptides with high hydrophobicity tended to have high pI. In contrast, gluC showed a negative correlation ($r = -0.30$ to -0.60) for the relationship between hydrophobicity and pI. Thus, the more hydrophobic peptides tended to have low pI, and the more hydrophilic peptides tended to have higher pI. Chymotrypsin was intermediate between trypsin and gluC with weak correlation ($r = 0.21$ to -0.28) between pI and hydrophobicity and with the slopes of the regression lines closer to zero. In the mass range 1000 to 2000, the trend was weakly negative, i.e. the chymotryptic peptides with the lowest pI were slightly more hydrophobic.

Discussion

The fractional mass is an inherent property of peptides, and the elements included in peptides have a different impact on the peptide's fractional mass. Hydrogen has the largest impact to push the fractional mass upwards. Sulfur, together with oxygen, pushes the fractional mass down. For the use of fractional mass to support interpretations of experimental data, relatively high accuracy (±10 ppm to 15 ppm, or better) is needed, but this can now be obtained with several types of MS instrument.

We found that the three proteases, trypsin, chymotrypsin and gluC, give similar, but not identical, regression lines for fractional mass vs integer mass of the peptides generated from 13,359 human protein sequences. The regression line for gluC peptides is slightly steeper than for trypsin or chymotrypsin. The differences in slope are related to the fact that the distributions of amino acid in the peptides are not random due to the cleavage specificities of the proteases, fully supporting the recent results of Wolski *et al*.[22] This is also the explanation for the intercept of the regression lines to vary from positive (trypsin) to negative (gluC). In fact, the C-terminal amino acid will have more influence the shorter the peptides are, and thereby tend to increase the deviation from the origin when compared to longer peptides. Certain types of missed cleavages, for example, such as two subsequent arginines or lysines, or the combination of arginine and lysine, which are sometimes encountered in a tryptic digest, would affect the fractional mass of the individual peptide. However, in our experience most tryptic peptides identified in an experimental setting are completely digested and we have therefore chosen to concentrate on peptides with no missed cleavages, although we acknowledge that missed cleavages and unexpected cleavages are more common for several other proteases, for example, chymotrypsin.

There is a high correlation between fractional mass and hydrophobicity for tryptic peptides when a narrow mass range is considered,[9] but as far as we know, this has not been investigated for other proteases. GluC-generated peptides have properties that are quite different from tryptic peptides. Since there is no restriction on the number of arginines or lysines present in a peptide, there is a stronger tendency for peptides with high fractional mass to have a high pI. These peptides would also have many charges, making them less hydrophobic. Thus, the regression lines for hydrophobicity and fractional mass tend to have a slope close to zero. On the other hand, peptides with high fractional mass could also be constituted by mainly hydrophobic amino acids, but this population of high fractional mass peptides is small compared to the population of high fractional mass/high pI peptides [see Figure 3(c)]. Overall, this would make peptides with low pI more hydrophobic than peptides with high pI. As chymotrypsin mainly cleaves C-terminal to the hydrophobic amino acids, chymotryptic peptides will not become very hydrophobic and the regression lines for hydrophobicity and fractional mass have a slope close to zero. On the other hand, there is no restriction on the number of acidic or basic amino acids in the peptides, resulting in a strong positive correlation between fractional mass and pI, while this correlation is weaker for hydrophobicity vs fractional mass.

The large differential fractional mass of certain elements has been utilized by the introduction of a mass defect tag to enhance the identification of peptides and proteins by tandem mass spectrometry (MS/MS). With the newer, and more accurate, MS/MS instruments it could become possible to ascribe fragments into N-terminal (a, b, c) or C-terminal (x, y, z) type by taking fractional mass into consideration without the incorporation of a mass defect tag. The speed of *de novo* sequencing algorithms (for example, References 15 and 26) may be enhanced, as they can allocate computing power to the more likely possibilities by using fractional mass as an additional filter.

The location of the fractional mass of an unidentified peak relative to the average fractional mass may give valuable hints on the amino acid composition of the peptide, which can then be compared with the amino acid sequence of the protein. For example, cysteine-containing peptides (also when alkylated by iodoacetamide or iodoacetic acid) have a strong tendency to locate themselves to the lower half of the fractional mass distribution. The interpretation of the peptide's amino acid content can be combined with searches for unexpected cleavages.[19,27] Many proteins are post-translationally modified, and one of the major tasks of proteomic studies is to identify these modifications and their positions. As has been noted before,[7,28] certain modifications give distinct contributions to the peptide's fractional mass. This may be employed in more aim-directed searches for phosphopeptides, which would have a tendency to locate themselves in the lower half of the fractional mass distribution.[28] Other modifications may also significantly influence the fractional mass of the peptide. For example, lipid modifications will increase the fractional mass.

In summary, some properties of the peptides can be directly explained by the cleavage specificity of the protease used. The

increasing accuracy of the MS instruments can, therefore, be used to exploit the information embedded in the fractional mass and give valuable hints on the amino acid composition of unidentified peptide peaks in a PMF.

Acknowledgments

H.B. is supported by the University of Bergen. V.C. was supported by the Norwegian Cancer Society.

References

1. W.J. Henzel, T.M. Billeci, J.T. Stults, S.C. Wong, C. Grimley and C. Watanabe, "Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases", *Proc. Natl. Acad. Sci. USA* **90,** 5011–5505 (1993). doi: 10.1073/pnas.90.11.5011
2. M. Mann, P. Højrup and P. Roepstorff, "Use of mass spectrometric molecular weight information to identify proteins in sequence databases", *Biol. Mass Spectrom.* **22,** 338–345 (1993). doi: 10.1002/bms.1200220605
3. P.James, M. Quadroni, E. Carafoli and G. Gonnet, "Protein identification by mass profile fingerprinting", *Biol. Mass Spectrom.* **22,** 338–345 (1993). doi: 10.1006/bbrc.1993.2009
4. M. Mann, "Useful tables of possible and probable petide masses", *43rd ASMS Conference on Mass Spectrometry and Allied Topics*, Atlanta, GA, USA, May 21–26 (1995).
5. S. Gay, P.A. Binz, D.F. Hochstrasser and R.D. Appel, "Modelling peptide mass fingerprinting data using the atomic composition of peptides", *Electrophoresis* **20,** 3527–3534 (1999). doi: 10.1002/(SICI)1522-2683(19991201)20:18<3527::AID-ELPS3527>3.0.CO;2-9
6. A. Wool and Z. Smilansky, "Precalibration of matrix-assisted laser desorption/ionization-time-of-flight spectra for peptide mass fingerprinting", *Proteomics* **2,** 1365–1373 (2002).
7. W.D. Lehmann, A. Bohne and C.W. von der Lieth, "The information encrypted in accurate peptide masses—improved protein identification and assistance in glycopeptide identification andcharacterization", *J. Mass Spectrom.* **35,** 1335–1241 (2000).
8. F. Schmidt, M. Schmid, P.R. Jungblut, J. Mattow, A. Facius and K.P. Pleissner, "Iterative data analysis is the key for exhaustive analysis of peptide mass fingerprints from proteins separated by two-dimensional electrophoresis", *J. Am. Soc. Mass Spectrom.* **14,** 943–956 (2003). doi: 10.1016/S1044-0305(03)00345-3
9. E.D. Dodds, H.J. An, P.J. Hagerman and C.B. Lebrilla, "Enhanced peptide mass fingerprinting through high mass accuracy: Exclusion of non-peptide signals based on residual mass", *J. Proteome Res.* **5,** 1195–1203 (2006). doi: 10.1021/pr050486o
10. J.A. Karty, M.M. Ireland, Y.V. Brun and J.P. Reilly, "Artifacts and unassigned masses encountered in peptide mass mapping", *J. Chromatogr. B: Analyt. Technol. Biomed. Life Sci.* **782,** 363–383 (2002). doi: 10.1016/S1570-0232(02)00550-0
11. W.A. Harris, D.J. Janecki and J.P. Reilly, "Use of matrix clusters and trypsin autolysis fragments as mass calibrants in matrix-assisted laser desorption/ionization time-of-flight mass spectrometry", *Rapid Commun. Mass Spectrom.* **16,** 1714–1722 (2002). doi: 10.1002/rcm.775
12. J. Samuelsson, D. Dalevi, F. Levander and T. Rognvaldsson, "Modular, scriptable and automated analysis tools for high-throughput peptide mass fingerprinting", *Bioinformatics* **20,** 3628–3635 (2004). doi: 10.1093/bioinformatics/bth460
13. R.A. Zubarev, P. Håkansson and B. Sundqvist, "Accuracy requirements for peptide characterization by monoisotopic molecular mass measurements", *Anal. Chem.* **68,** 4060–4063 (1996). doi: 10.1021/ac9604651
14. B. Spengler, "*De novo* sequencing, peptide composition analysis and composition-based sequencing: a new strategy employing accurate mass determination by Fourier transform ion cyclotron resonance mass spectrometry", *J. Am. Soc. Mass Spectrom.* **15,** 704–715 (2004). doi: 10.1016/j.jasms.2004.01.007
15. B. Spengler, "Accurate mass as a bioinformatic parameter in data-to-knowledge conversion: Fourier transform ion cyclotron resonance mass spectrometry for peptide de novo sequencing", *Eur. J. Mass Spectrom.* **13,** 83–87 (2007). doi: 10.1255/ejms.840
16. A.A. Schlosser and W.D. Lehmann, "Patchwork peptide sequencing: extraction of sequence information from accurate mass data of peptide tandem mass spectra recorded at high resolution", *Proteomics* **2,** 524–533 (2002). doi: 10.1002/1615-9861(200205)2:5<524::AID-PROT524>3.0.CO;2-O
17. Swiss-Prot: http://au.expasy.org/sprot/
18. Swissknife: http://swissknife.sourceforge.net/docs/
19. H. Barsnes, S.O. Mikalsen and I. Eidhammer, "MassSorter: a tool for administrating and analyzing data from mass spectrometry experiments on proteins with known amino acid sequences", *BMC Bioinformatics* **7(42),** (2006).
20. pI_tool: http://us.expasy.org/tools/pi_tool.html
21. J. Kyte and R.F. Doolittle, "A simple method for displaying the hydropathic character of a protein", *J. Mol. Biol.* **157,** 105–132 (1982). doi: 10.1016/0022-2836(82)90515-0
22. W.E. Wolski, M. Farrow, A.K. Emde, H. Lehrach, M. Lalowski and K. Reinert, "Analytical model of peptide mass cluster centres with applications", *Proteome Sci.* **4(18),** (2006). doi: 10.1186/1477-5956-4-18
23. B.J. Cargile, D.L. Talley and J.L. Stephenson, Jr, "Immobilized pH gradients as a first dimension in shotgun proteomics and analysis of the accuracy of pI predictability of peptides", *Electrophoresis* **25,** 936–945 (2004). doi: 10.1002/elps.200305722

24. B.J. Cargile, J.R. Sevinsky, A.S. Essader, J.L. Stephenson, Jr and J.L. Bundy, "Immobilized pH gradient isoelectric focusing as a first-dimension separation in shotgun proteomics", *J. Biomol. Tech.* **16,** 181–189 (2005).

25. M.P. Hall, S. Ashrafi, I. Obegi, R. Petesch, J.N. Peterson and L.V. Schneider, "'Mass defect' tags for biomolecular mass spectrometry", *J. Mass Spectrom.* **38,** 809–816 (2003). doi: 10.1002/jms.493

26. A.M. Frank, M.M. Savitski, M.L. Nielsen, R.A. Zubarev and P.A. Pevzner, "*De novo* peptide sequencing and identification with precision mass spectrometry", *J. Proteome Res.* **6,** 114–123 (2007). doi: 10.1021/pr060271u

27. FindPept: http://au.expasy.org/tools/findpept.html

28. C. Bruce, M.A. Shifman, P. Miller and E.E. Gulcicek, "Probabilistic enrichment of phosphopeptides by their mass defect", *Anal. Chem.* **78,** 4374–82 (2006). doi: 10.1021/ac060046w