Investigating the structural factors governing efficiency of ionisation by nano-electrospray of tryptic peptides

***Rebecca J. Mackenzie* [1], *Stephen W. Holman* [1], *Claire E. Eyers* [1]**

[1] Michael Barber Centre for Mass Spectrometry, School of Chemistry, Manchester, Institute of Biotechnology, The University of Manchester, 131 Princess Street, Manchester, M1 7DN, UK

Abstract

Determining the extent to which specific physico-chemical parameters govern the ionisation efficiency and thus the detection of peptides is yet to be resolved. Using a set of 63 QconCAT proteins, which upon proteolysis yielded 63 stoichiometric mixtures of tryptic quantification (or Q) peptides (2658 peptides in total), calculated parameter values (*e.g.* mean positive charge), were compared to normalised peak areas for each peptide to evaluate the relationship between parameter and peptide detection. Our results confirm that no single parameter governs peptide ionisation efficiency by nano-electrospray, rather a number of parameters including hydrophobicity, secondary structure and polarity appear to regulate ionisation efficiency.

Introduction

Advances in mass spectrometry have enabled the extensive study of biological systems through identification, characterisation and quantification of their protein constituents[1]. Absolute protein quantification is primarily carried out at the peptide level, using the ratio between unknown amounts of native peptide to known quantities of an identical isotope-labelled reference, to quantify the analyte. The requirement for such standards has resulted in the production of isotope-labelled quantification peptides (Q peptides), synthesised *de novo via* chemical methods, typically referred to as AQUA peptides[2]. Alternatively, artificial genes can be designed *de novo* to mediate synthesis of novel proteins (QconCAT proteins) that are assemblies of the signature Q peptides[3]. To overcome the requirement for such reference standards there needs to be better understanding of the mechanisms of (nano-) electrospray ionisation and crucially, the physico-chemical parameters that regulate signal intensity.

The observation of a peptide in an LC-MS experiment, *i.e.* its ionisation efficiency is governed by various parameters, including mass, charge, isoelectric point, hydrophobicity and features relating to secondary structure. These, along with a number of other features, were shown to be important for the *in silico* prediction of peptide 'flyers' using a combination of computational algorithms[4], providing an insight into properties that govern detectability. However, the specific physico-chemical features that most heavily influence peptide ionisation efficiency by electrospray remain poorly understood. We have extended our previous studies, making use of a quality controlled stoichiometric test set of over 1600 tryptic peptides, derived from 63 QconCAT constructs, to further investigate those peptide features that govern mass spectrometric signal response. This collection of peptides is unique in that it provides a large test set that can be used to assess ionisation efficiency independent of concentration.

Methods

Each of 63 isotopically labelled QconCAT proteins ($^{13}C_6$ [Arg/Lys]; ~0.05µg µL^{-1}, in 25 mM ammonium bicarbonate) were reduced and alkylated prior to overnight digestion with 2% (w/w) trypsin (Sigma-Aldrich, Poole, UK) resulting in a final protein concentration of 100 fmol µL^{-}1. 1 µL of each digested QconCAT was then analysed in triplicate by liquid chromatography-mass spectrometry (LC-MS) using a nanoACQUITY ultra-performance liquid chromatograph (Waters Ltd., Elstree, UK) and a Synapt HDMS quadrupole-time-of-flight (Q-

ToF) mass spectrometer (Waters Ltd., Elstree, UK). Peptides were loaded onto a trapping column (Waters, Symmetry C_{18}, 180 µm x 20 mm, 5 µm), with partial loop injection and trapped for 3 min (5 µL min^{-1}) with 99.9% mobile phase A, 0.1% mobile phase B (v/v) (A = 0.1% formic acid in water, B = 0.1% formic acid in acetonitrile). Peptides were resolved on an analytical column (Waters, nanoACQUITY UPLC^{TM} HSS T3, 75 µm x 150 µm, 1.8 µm) using a gradient of 97% A, 3% B (v/v) to 60% A, 40% B (v/v) over 10 min at a flow rate of 300 nL min^{-1}. The column was washed with 5% A, 95% B and then re-equilibrated to starting conditions. A lock mass solution of 500 fmol µL^{-1} glu-fibrinopeptide B was infused into the nESI source at a flow rate of 300 nL min^{-1}.

The column effluent was introduced into a nESI source fitted with a PicoTip emitter (New Objective, Woburn, MA, USA). The ionisation source polarity was set to positive and typical operating conditions were: capillary voltage, 2.5 kV; cone voltage, 25 V; extraction cone, 4 V; source temperature, 70 ^{o}C. Data acquisitions were performed using an MS^E experiment[5]. The survey scan was performed between *m*/*z* 50-2000 with a scan time of 300 msec and a trap cell collision energy (CE) of 6 eV. Product ion spectra were acquired using the same *m*/*z* range and scan time, but with a trap cell CE starting at 15 eV and ramping to 40 eV. The lock mass was sampled every 30 sec using a trap cell CE of 6 eV and a cone voltage that facilitated a detector response of between 100-200 counts per second.

Following acquisition, data was processed using ProteinLynx Global Server (PLGS) v.2.5.2 (Waters, Elstree, UK) to identify the detected peptides. Product ion spectra were searched against an in-house generated database of QconCAT proteins. Fixed modifications of carbamidomethyl Cys and $^{13}C_6$ [Arg/Lys] and variable modification of oxidised Met were specified. One trypsin missed cleavage was allowed. The raw data was imported into Skyline v.1.3.0.3871[6] for visual assessment and comparison with PLGS. Data was further analysed where necessary using IDCalc v.0.3 (http://proteome.gs.washington.edu/software/IDCalc/) and raw data reviewed using MassLynx v.4.1 (Waters). Peptide sequences observed as a missed cleavage pair were removed from the dataset, whilst those that were not observed in any form were assigned a peak area of 0. For each QconCAT protein, the peak area as determined by Skyline for the most abundant precursor ion charge state for each Q peptide was normalised relative to the doubly-protonated heavy glu-fibrinopeptide B precursor ion before being averaged over the three technical replicates, thereby normalising for differences in column loading between samples.

An in-house developed program was used to determine values for 1180 physico-chemical properties for each peptide. The peak areas were then correlated with the parameter values in the statistical package R (http://www.r-project.org/) using Pearson correlation, such that the relationship between all peptide peak areas and all values for a specific parameter was determined. For peptides with similar peak intensities for different precursor ion charge states, both peak areas and charges were included. Absolute parameter coefficients were then sorted in numerical order to determine which parameters had the strongest relationship with peak area.

Results and Discussion

From a sample of 63 QconCATs, 2658 tryptic Q peptides were theoretically possible from an *in silico* digest. Of these 2658 theoretical peptides, 1684 were observed as limit peptides, 538 were involved in missed cleavages and a further 436 were not observed in these nLC-MS^E experiments and thus were assigned peak areas of 0.

The prevalence of missed cleavages in the dataset resulted in the removal of ~20% of the total peptides. Not unexpectedly, these peptides contained acidic residues either upstream or downstream of the cleavage site and fit current models regarding motifs that increase the propensity for missed cleavage. The presence of such acidic residues (glutamate and aspartate), in close proximity (positions P1' and P2') to the cleavage site has been demonstrated to inhibit the generation of limit peptides[7]. These residues act to occupy the catalytic basic side chain of trypsin through formation of a salt bridge, competing successfully against the arginine and lysine residue in the peptide, resulting in promotion of missed cleavages[8]. Q peptides which were allocated a peak area of 0 produced chromatograms in which the precursor ions and y ions either did not follow a similar elution profile, had mass spectra with incorrect isotopic ratios or had peaks with low signal-to-background ratio and so the highest intensity signal for precursor ions (referred to as 'peak' in this report) was not successfully identified (Figure 1).

Unlike PLGS, Skyline attempts to assign a peak to a peptide sequence, providing an i-dot-product (i-dotp) value as a measure of likelihood of correct assignment. This can be problematic, as peptides that have not generated any LC-MS data are assigned to peaks in Skyline, producing erroneous data. In addition, peptides that have ionised may be assigned to another peptide's peak. PLGS, on the other hand, sometimes does not report identification of some peptides annotated by Skyline, often due to insufficient product ion information to support identification. For further evaluation of the quality of peptide spectra, IDCalc was used to compare expected precursor ion isotopic ratios. If the observed isotope ratio did not match the expected ratio, it was deemed low quality. As a final point of clarification, the raw data was visualised using MassLynx. Using these methods of data interrogation, 1684 Q peptides were deemed suitable. For these, peak area (and retention time) was determined using Skyline.

Of those peptides assigned a peak area of 0, 7% were deemed not to be detectable in an LC-MS experiment by CONSeQuence. Of the remaining, many did not fragment sufficiently to yield a positive identification. Peptide detectability in such experiments is a measure of both its ability to be separated by reversed-phase chromatography and its propensity to ionise. It is therefore possible that peptides that were not detected did not chromatograph well, being either too hydrophilic and therefore not binding to the C18 resin, or too hydrophobic and not eluting under gradient conditions. Comparing the sum buriability[9] (an indication of hydrophobicity) between positively identified peptides and those that were assigned a peak area of 0, buriability factors of 12.72 and 13.12 resp were calculated, indicating that in general, non-identified peptides were more hydrophobic[9]. In addition, sum buriability values were calculated in the extremes of the range (<8 and >18) for the non-identified peptides. The size of the negative dataset precludes any definitive conclusion regarding the effect of hydrophobicity on non-identification of peptides, although a trend is apparent.

Following peak assignment, the peak area for each peptide was correlated against each of the 1180 different peptide features[4] using Pearson correlation in R. Values did not exceed 0.3000 (irrespective of relationship) (Figure 2). Interestingly, in these stoichiometrically controlled studies, for the top 50 parameters (20 of which are shown in Table 1), none of these were identical to the top 50 ranked parameters governing peptide detectability identified by Eyers *et al*[4]. Nevertheless, a number of the top 50 parameters identified between these two studies are related: for instance, the parameter describing the propensity to be 'buried inside' observed in these studies[10] is related to 'buriability' defined in our previous analysis[9]. Many of the parameters in Table 1 are related, for example four different methods (A, B, C, D) for calculating 'optimised relevant partition energies'[11] appear in the top 20 features, having a positive influence on ionisation efficiency[12]. These methods for optimised relevant partition energies differ due to the contact energies taken into consideration when calculating the relevant partition energies. For example, Method B corresponds to cases in which interactions among residues consist only of pairwise contact energies whilst Method D corresponds to total interaction energies, consisting of contact energies, repulsive packing energies and secondary structure energies[11]. Although these methods share similar energies and thus can be grouped, they are not identical and so can be listed as four separate parameters. Other parameters may be grouped into a single parameter, including the hydrophobicity index[13] and the hydrophobicity scales[14]. Such grouping was carried out in the previous study[4], demonstrating a possibility as to why the top parameter lists differ.

The maximum Pearson correlation observed in these studies was 0.3, which is significantly lower than the value of 0.7 observed by Eyers *et al*[4] during the development of CONSeQuence. This suggests that whilst parameters such as buriability and hydrophobicity influence peptide detection, it is a cumulative effect of many parameters that enable them to efficiently ionise. Despite this cumulative effect, some properties appear to have a stronger relationship with peak area, having a higher Pearson correlation value. Those which do appear to have a stronger positive effect include the optimized relative partition energies and hydrophobicity index. The linker index provides the strongest negative effect, as it predicts whether a sequence is part of an interdomain protein region. A higher linker index indicates a lower level of secondary structure [15]. Furthermore it could be the case that not all parameters tested may contribute to efficient ionisation. Some parameters may have more of an effect than others, whilst others may have no effect at all. Additionally, some parameters may counter the effect of others, which Pearson correlation alone cannot determine. Various other statistical tests will be carried out to investigate these points further, for example, two-way ANOVA. However, this requires automation due to the large number of variables involved.

Conclusion

Whilst Eyers *et al.* were able to predict detectable peptides using CONSeQuence[4], giving insight into which parameters determine detectability (related to both LC and ESI), due to the nature of the sample set used for training, the differences in amount, the extent to which the peptides could be detected required further investigation. As CONSeQuence only determined absolute rather than relative ability for detection of a peptide sequence in an LC-MS experiment, the top physico-chemical parameters defined by Eyers *et al.* differ from those determined in this experiment.

Whilst we now have an insight into the extent to which various parameters affect the efficiency of ionisation, much more work is required to identify the interplay between physico-chemical properties that ultimately determine ionisation efficiency. Such work can be carried out on the acquired data set using statistical measures to determine relationships between parameters. However we aim to further improve the current dataset by both increasing the number of limit peptides and undertaking a concerted evaluation of a negative dataset.

Acknowledgements

The authors wish to thank members of the Michael Barber Centre for Mass Spectrometry; Dr Craig Lawless for the in-house program for determining parameter values, Victoria Harman and Rachel Watkins for provision of the QconCAT proteins; and the CoPY Project (at the Universities of Manchester and Liverpool). In addition, the authors acknowledge the British Mass Spectrometry Society and the Royal Society of Chemistry Northwest Analytical Division for funding.

References

1. R. Aebersold and M. Mann, *Nature*, 2003, **422**, 198-207.
2. S. A. Gerber, J. Rush, O. Stemman, M. W. Kirschner and S. P. Gygi, *Proceedings of National Academy of Science U S A*, 2003, **100**, 6940-6945.
3. J. Rivers, D. M. Simpson, D. H. Robertson, S. J. Gaskell and R. J. Beynon, *Molecular & Cellular Proteomics*, 2007, **6**, 1416-1427.
4. C. E. Eyers, C. Lawless, D. C. Wedge, K. W. Lau, S. J. Gaskell and S. J. Hubbard, *Molecular & Cellular Proteomics*, 2011, **10**, M110 003384.
5. R. S. Plumb, K. A. Johnson, P. Rainville, B. W. Smith, I. D. Wilson, J. M. Castro-Perez and J. K. Nicholson, *Rapid Communications in Mass Spectrometry*, 2006, **20**, 1989-1994.
6. B. MacLean, D. M. Tomazela, N. Shulman, M. Chambers, G. L. Finney, B. Frewen, R. Kern, D. L. Tabb, D. C. Liebler and M. J. MacCoss, *Bioinformatics*, 2010, **26**, 966-968.
7. C. Lawless and S. J. Hubbard, *OMICS*, 2012.
8. B. Thiede, S. Lamer, J. Mattow, F. Siejak, C. Dimmler, T. Rudel and P. R. Jungblut, *Rapid Communications in Mass Spectrometry*, 2000, **14**, 496-502.
9. H. Zhou and Y. Zhou, *Proteins*, 2004, **54**, 315-322.
10. D. H. Wertz and H. A. Scheraga, *Macromolecules*, 1978, **11**, 9-15.
11. S. Miyazawa and R. L. Jernigan, *Proteins*, 1999, **34**, 49-68.
12. N. Qian and T. J. Sejnowski, *Journal of Molecular Biology*, 1988, **202**, 865-884.
13. G. D. Fasman, *Prediction of Protein Structure and the Principles of Protein Conformation*, Plenum, New York, 1989.
14. P. K. Ponnuswamy, *Progress in Biophysics and Molecular Biology*, 1993, **59**, 57-103.
15. K. Bae, B. K. Mallick and C. G. Elsik, *Bioinformatics*, 2005, **21**, 2264-2270.

Table 1. Top 20 parameters influencing peptide ionisation efficiency as defined by Pearson correlation

	Parameter	R Value
	sum:BAEK050101 Linker index (Bae et al. 2005)	-0.2987
2	sum:GUYH850102 Apparent partition energies calculated from Wertz-Scheraga index (Guy 1985)	0.2942
3	sum:MIYS990102 Optimized relative partition energies - method A (Miyazawa-Jernigan 1999)	0.2896
4	sum:MIYS990101 Relative partition energies derived by the Bethe approximation (Miyazawa-Jernigan 1999)	0.2895
5	sum:MIYS990103 Optimized relative partition energies - method B (Miyazawa-Jernigan 1999)	0.2886
6	sum:MIYS990104 Optimized relative partition energies - method C (Miyazawa-Jernigan 1999)	0.288
7	sum:BASU050102 Interactivity scale obtained by maximizing the mean of correlation coefficient over single-domain globular proteins (Bastolla et al. 2005)	-0.2808
8	sum:FASG890101 Hydrophobicity index (Fasman 1989)	0.2792
9	sum:MIYS990105 Optimized relative partition energies - method D (Miyazawa-Jernigan 1999)	0.2724
10	sum:GUYH850101 Partition energy (Guy 1985)	0.2688
11	sum:NADH010104 Hydropathy scale based on self-information values in the two-state model (20% accessibility) (Naderi-Manesh et al. 2001)	-0.2678
12	sum:NADH010105 Hydropathy scale based on self-information values in the two-state model (25% accessibility) (Naderi-Manesh et al. 2001)	-0.2669
13	sum:NISK860101 14 A contact number (Nishikawa-Ooi 1986)	-0.2662
14	sum:BIOV880101 Information value for accessibility; average fraction 35% (Biou et al. 1988)	-0.2631
15	mean:WERD780101 Propensity to be buried inside (Wertz-Scheraga 1978)	-0.2614
16	mean:GUYH850102 Apparent partition energies calculated from Wertz-Scheraga index (Guy 1985)	0.2609
17	mean:BAEK050101 Linker index (Bae et al. 2005)	-0.2609
18	sum:PONP930101 Hydrophobicity scales (Ponnuswamy 1993)	-0.2599
19	sum:CORJ870103 PRIFT index (Cornette et al. 1987)	-0.2589
20	sum:NOZY710101 Transfer energy organic solvent/water (Nozaki-Tanford 1971)	-0.2576

Figure Legends

1.) Evaluation of spectral quality

a) MS/MS spectra and IDCalc calculated isotopic ratios for peptide A, which Skyline gave an i-dotp value of 0.80. The elution profiles between precursor and product ions are dissimilar whilst the isotopic ratios do not meet the calculated values, so this peptide was assigned a peak area of 0; b) MS/MS spectra and IDCalc calculated isotopic ratios for peptide B, which Skyline gave an i-dotp value of 0.99. The elution profiles share similar shape and retention time, whilst the isotopic ratios agree, concluding this peptide to be correctly assigned.

2.) Histogram of R values for all 1180 parameters

Calculating the correlation between all peak areas and an individual parameter gave an R value. These R values were less than 0.3000, with a higher number of parameters having a negative effect on peak area.