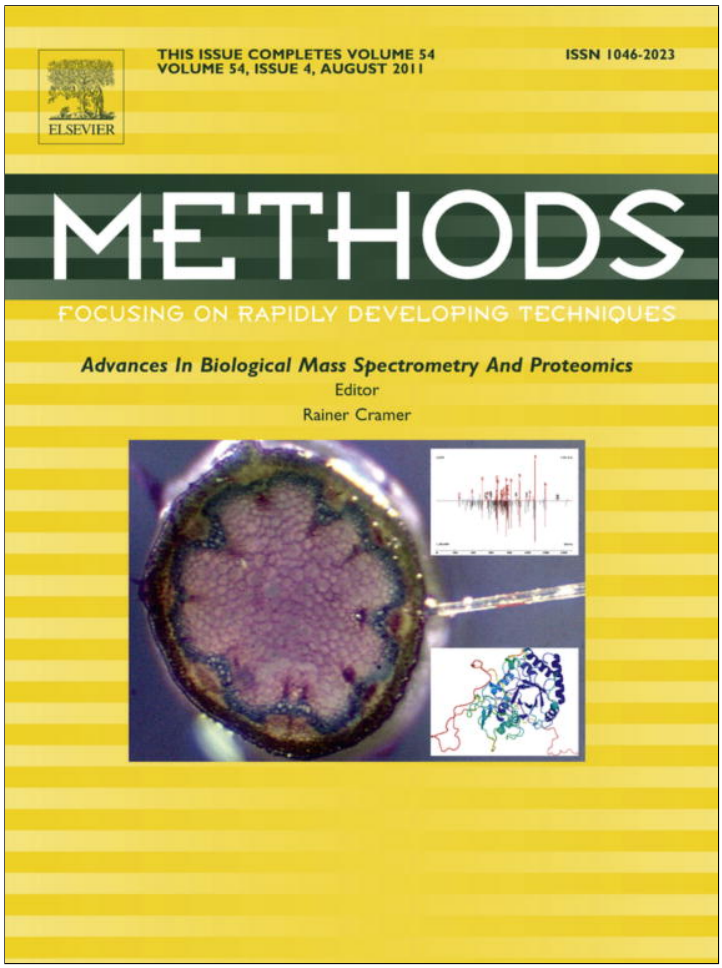

Methods 54 (2011) 351–360

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

The importance of the digest: Proteolysis and absolute quantification in proteomics

Philip Brownridge, Robert J. Beynon *

Protein Function Group, Institute of Integrative Biology, University of Liverpool, Liverpool L69 7ZB, UK

ARTICLE INFO

Article history:
Available online 6 June 2011

Keywords:
Quantitative proteomics
QconCAT
Proteolysis
Trypsin
Selected reaction monitoring
Quantotypic peptides
Endopeptidase LysC
Label-free quantification
Stable isotope labelling

ABSTRACT

Virtually all mass spectrometric-based methods for quantitative proteomics are at the peptide level, whether label-mediated or label-free. Absolute quantification in particular is based on the measurement of limit peptides, defined as those peptides that cannot be further fragmented by the protease in use. Complete release of analyte and (stable isotope labelled) standard ensures that the most reliable quantification data are recovered, especially when the standard peptides are in a different primary sequence context, such as sometimes occurs in the QconCAT methodology. Moreover, in label-free methods, incomplete digestion would diminish the ion current attributable to limit peptides and lead to artifactually low quantification data. It follows that an essential requirement for peptide-based absolute quantification in proteomics is complete and consistent proteolysis to limit peptides. In this paper we describe strategies to assess completeness of proteolysis and discuss the potential for variance in digestion efficiency to compromise the ensuing quantification data. We examine the potential for kinetically favoured routes of proteolysis, particularly at the last stages of the digestion, to direct products into 'dead-end' mis-cleaved products.

1. Introduction

In the 1940's, Linderstrøm-Lang studied the action of proteases on proteins. He proposed two different mechanisms; an 'all or none' process, whereby a protease bound to a substrate molecule and remained associated until the protein was fully digested, and a 'zipper' process, whereby a protease interacted with intact substrate and partially degraded fragments, until proteolysis was complete [1]. The primary difference between the two processes was that in the former, there could be no degradation intermediates free in the digestion reaction whereas the zipper mechanisms could release partially cleaved products. We now know that the all-or-none mechanism does not operate (other than in the confines of the 20S proteasomal core), and that degradation intermediates are therefore not only likely but obligatory for simple endopeptidases.

When a protein undergoes an initial proteolytic event, the products can become more or less susceptible to further proteolysis. For example, the proteolytic action of enteropeptidase on trypsinogen produces active trypsin by virtue of the loss of an N-terminal hexapeptide. The activated enzyme is less likely to undergo further proteolysis by enteropeptidase – if this were not the case; the active enzyme would be degraded more rapidly and would not persist. Alternatively, a protein can be destabilised by the initial proteolytic cleavage, such that the products are more rapidly cleaved into multiple further products. A feature of the latter behaviour is that the route of digestion might not follow the same pathway for each protein molecule, and thus, a large number of discrete, partially digested species are generated. It is only as the sequential proteolytic reactions reach completion that the different pathways converge to the same products (Fig. 1). When all peptide bonds that **can** be cleaved **have** been cleaved, the resultant set of peptides are referred to as 'limit peptides'; peptides that lack any further endoproteolytic sites compatible with the endopeptidase being used.

Despite the development of top-down analytical approaches, most proteomics workflows require a proteolytic step prior to mass spectrometric analysis of the peptides generated by the hydrolytic reaction. In most instances, the endopeptidase that is used is trypsin, reflecting the very restricted specificity of this enzyme (Arg-X, Lys-X, and under normal circumstances, zero or low frequency cleavage at Arg-Pro, Lys-Pro) and the fact that most products from a tryptic digest have a minimum of two protonatable sites (the N-α amino group and the C-terminal basic residue) and thus generate $[M+2H]^{2+}$ ions, enhancing the generation and enhancement of gas phase fragmentation products. In many proteomes the residues arginine and lysine are each present at about 5% of the amino acids, making a tryptic fragment approximately 10–15 amino acids residues long. Assuming complete fragmentation, the limit peptides that are detectable are usually between 1000 (about eight amino acids) and 3000 Da (about 25 amino acids), optimally aligned to the m/z range of the mass analysers used in mass spectrometers that feature in proteomics studies.

* Corresponding author.
E-mail address: r.beynon@liv.ac.uk (R.J. Beynon).

1046-2023/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ymeth.2011.05.005

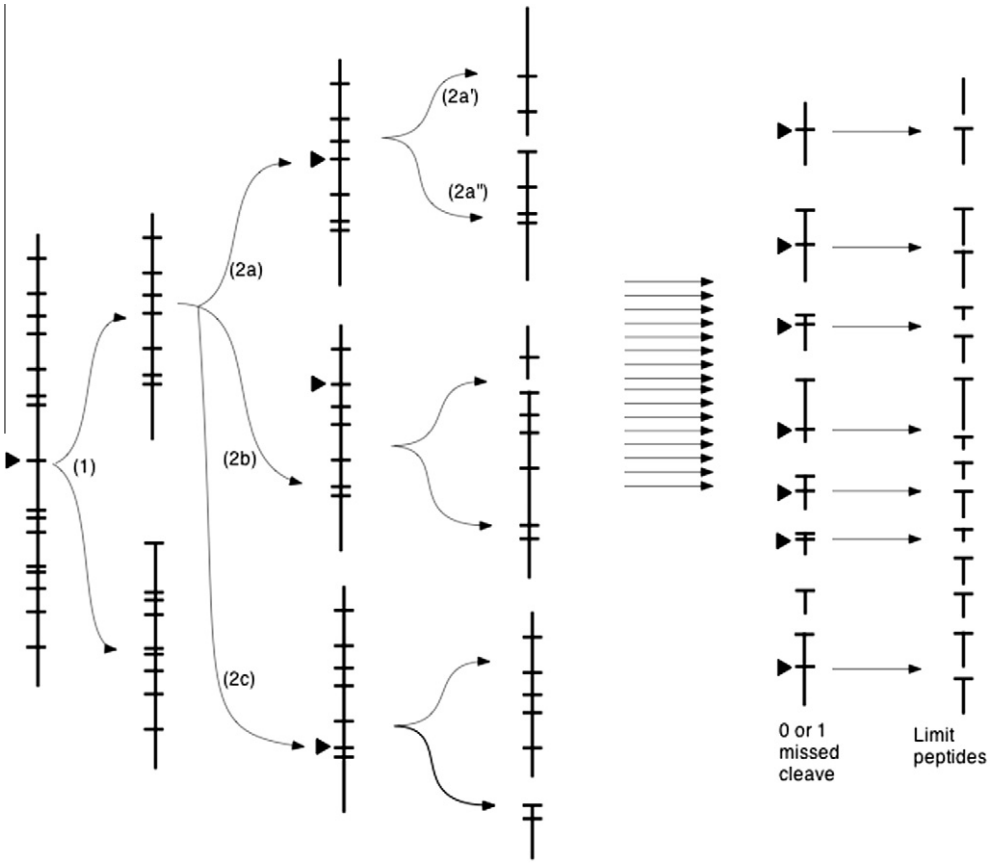

Fig. 1. Routes of proteolysis of a protein. In the absence of higher order structural factors that can modify the propensity of sites to be digested, the conversion of an intact protein to limit peptides can take many different routes; the relative occupancy of such routes is a consequence of the intrinsic digestibility of each scissile bond. Eventually, multiple pathways converge to oligopeptides that are defined as 'mis-cleaved peptides'. Completeness of digestion can be assessed by monitoring the ratio of mis-cleaved peptides and limit peptides.

Although proteolysis of an entire proteome is often predicated on the complete hydrolysis of all proteins to limit peptides, many protein identification strategies are tolerant to a small number of mis-cleavages (typically one or two), which might even enhance the strength of the identification, since a mis-cleaved product restores some of the lost connectivity that is inherent in a set of limit peptides – in a fully proteolysed proteome we do not know which peptides are 'adjacent' to each other. The gain in identifiability is more critical in peptide mass fingerprinting, because the only piece of information obtained from the peptide is the mass whereas in tandem mass spectrometry further information is gained from each peptide according to sequence specific fragmentation.

Although it may be possible to optimise complete proteolysis for a single protein, a proteome offers a large and complex reaction space. Residues at least three positions distal to the cleavage site can affect proteolysis, predominantly through changes in the affinity of the endopeptidase for the substrate [2]. This creates a large number of possible (approximately $20^6 = 64$ million) different cleavage sites, although in practice only a subset of these are evident in any proteome; for example, the *Saccharomyces cerevisiae* proteome has approximately 250,000 tryptic sites. Many of the tryptic sites will be efficiently and completely cleaved, but some, for example, those with acidic residues C-terminal to the scissile bond in positions P1′ and P2′ (nomenclature of Schechter and Berger [3]), will be slow to hydrolyse and therefore difficult to digest to completion [4].

Although mis-cleaved products can sometimes enhance the quality of an identification workflow, there are circumstances in which they can compromise quantitative proteomics. Peptide-level quantification can be conducted by label-mediated methods or label-free approaches. In label-mediated methods, a differentially stable isotope labelled standard peptide [5] of known amount is co-analysed with the analyte, and the ratio of the analyte to standard reveals the abundance of the analyte. Relative quantification (whether label free or isotope coded, such as is obtained with metabolically labelled samples) may be more tolerant to incomplete proteolysis, provided that it can be assumed that the labelled and unlabelled proteins undergo the same extent of proteolysis. For absolute quantification, whether using stable isotope labelled chemically synthesized peptides (AQUA peptides) or peptides derived from hydrolysis of a protein standard (QconCAT or PSAQ) it is necessary to compare the analyte with a standard, usually at the peptide level, in assays wherein the quantities of one or more (tryptic) peptides are considered to be formally representative of the quantity of the parent protein. Complete proteolysis of analyte (AQUA) or analyte and standard (PSAQ, QconCAT) is thus far more critical in quantification workflows than it is in discovery workflows. Completeness of proteolysis is also important in label-free methods that make use of the number of tryptic fragments observable (spectral counting) or the inherent intensity of the mass spectrometric signal for one or more peptides [6,7]. Implicit in either of these approaches is that the optimal data will be obtained if the analyte (and in some instances, standard) signal is delivered by limit peptides.

The requirement for complete proteolysis is never more important than in QconCAT quantification workflows [8–11]. QconCATs are artificial proteins that are concatenated tryptic peptides from a large number of different analyte proteins, typically two peptides for each protein. The gene that would direct the synthesis of the QconCAT is synthesized *de novo*, and expressed heterologously in bacteria, in stable isotope labelled media. Once purified, a known

amount of the QconCAT is mixed with the analyte proteins and co-proteolysed, releasing mixtures of heavy and light standard and analyte tryptic peptides that can then be analysed by mass spectrometry. Since the amount of the standard is known, the analyte peptide, and by implication, protein, can be quantified by virtue of the relative intensities or ion chromatograms of the isotopologues. Because the tryptic fragments in the QconCAT are adjacent to other standard peptides, the primary sequence context of the standard and analyte may differ, and proteolytic excision of the peptide may occur at a different rate in the QconCAT than in the analyte. Thus, in a QconCAT workflow, it is essential that the digestion of both analyte and standard are complete [12]. This is an implicit requirement in the QconCAT protocol, since to be effective for quantification, both hydrolytic reactions must go to completion. The requirement is therefore to identify reaction conditions that guarantee complete digestion.

2. Materials and methods

2.1. Unless otherwise stated all chemicals were supplied by Sigma (Poole, UK)

2.1.1. Mis-cleaved peptide analysis

Four 160 μL biological replicate samples containing 100 μg of broken yeast preparation were digested with trypsin. The proteins were denatured with 10 μL of 1% (w/v) RapiGest™ (Waters) in 25 mM ammonium bicarbonate and incubation at 80 °C for 10 min. The sample was reduced (addition of 10 μL of 60 mM DTT and 10 min incubation at 65 °C) and alkylated (addition of 10 μL of 180 mM iodoacetamide and incubation at room temperature for 30 min in the dark). Trypsin (Roche Diagnostics Ltd., West Sussex, UK) was reconstituted in 50 mM acetic acid to a concentration of 0.2 μg/μL. Digestion was performed by the addition of 20 μL of trypsin to the sample followed by incubation at 37 °C. After 4.5 h 2.5 μL of 0.1 M hydrochloric acid and an additional 10 μL of trypsin was added and incubated overnight. Rapigest™ was removed by centrifugation following sample acidification (2 μL of formic acid and incubation at 37 °C for 45 min).

2.1.2. Time-course digest

A 320 μL sample containing 625 ng/μL yeast lysate, 6.25 fmol/μL QconCAT CC1 (targeting chaperone proteins) and 250 mM NaCl was prepared for the time-course digestion. The proteins were denatured using 20 μL of 1% (w/v) RapiGest™ (Waters, Manchester, UK) in 25 mM ammonium bicarbonate followed by incubation at 80 °C for 10 min. The sample was reduced (addition of 20 μL of 60 mM dithiothreitol and incubation at 65 °C for 10 min) and alkylated (addition of 20 μL of 180 mM iodoacetamide and incubation at room temperature for 30 min in the dark). Trypsin (Roche Diagnostics Ltd., West Sussex, UK) was reconstituted in 50 mM acetic acid to a concentration of 0.2 μg/μL. Digestion was performed by the addition of 20 μL of trypsin to the sample followed by incubation at 37 °C. At 0 s, 30 s, 1 min, 2 min, 5 min, 10 min, 15 min, 30 min, 1 h, 2 h and 4 h, a 10 μL portion was removed and mixed with 10 μL of 5% trifluoroacetic acid and refrigerated to terminate proteolysis. After 4.5 h 3.6 μL of 0.1 M hydrochloric acid and an additional 20 μL of trypsin was added and additional time points at 5 min, 10 min and 1 h were sampled following the enzyme top-up. After a final overnight time point had been sampled, all time point samples were incubated at 37 °C and centrifuged to remove the Rapigest™.

2.1.3. LysC–trypsin "double-digestion"

A 160 μL mixture of five $^{13}C_6$ lysine and $^{13}C_6$ arginine isotopically-labelled QconCAT proteins at approximately 100 fmol/μL against a background of 100 ng/μL yeast lysate was prepared. The proteins were denatured with 10 μL of 1% (w/v) RapiGest™ (Waters) in 25 mM ammonium bicarbonate and incubation at 80 °C for 10 min. The sample was reduced (addition of 10 μL of 60 mM DTT and 10 min incubation at 65 °C) and alkylated (addition of 10 μL of 180 mM iodoacetamide and incubation at room temperature for 30 min in the dark). Endoprotease Lys-C (Roche) was reconstituted in 50 mM acetic acid to a concentration of 0.1 μg/μL. Digestion was performed by the addition of 10 μL of Lys-C to the sample followed by incubation at 37 °C. A sample of this digest was taken and analysed by SDS–PAGE to confirm protein digestion. Trypsin (Roche) was added (10 μL of 0.2 μg/μL trypsin in 50 mM acetic acid) and the sample incubated overnight at 37 °C. At timepoints of 0 min, 30 s, 1 min, 2 min, 5 min, 10 min, 15 min, 30 min, 1 h, 2 h, 4 h, 8 h and overnight, 5 μL of sample was removed and mixed with 5 μL of 5% trifluoroacetic acid and refrigerated to terminate proteolysis. After final overnight analysis all samples were incubated at 37 °C and centrifuged to remove Rapigest™.

2.1.4. Mass spectrometry

All samples were analysed by LC–MS using a nanoAcquity UPLC™ system (Waters MS Technologies, Manchester, UK). The 1 μL sample was injected onto the trapping column (Waters, C18, 180 μm × 20 mm), using partial loop injection, for 3 min at a flow rate of 5 μL/min with 0.1% (v/v) formic acid. The sample was resolved on the analytical column (Waters, nanoACQUITY UPLC™ BEH C18 75 μm × 150 mm 1.7 μm column) using a gradient of 97% A (0.1% (v/v) formic acid) 3% B (99.9% acetonitrile 0.1% (v/v) formic acid) to 60% A 40% B over 30 min at a flow rate of 300 nL/min. For the time-course study the nanoAcquity UPLC™ was coupled to a Xevo™ TQ triple quadrupolar mass spectrometer (Waters) operated in scheduled SRM mode with Q1 and Q3 operating at unit resolution. Each peptide (both isotopic variants) was targeted by two transitions and the program set to acquire 15 datapoints over a 15 s chromatographic peak. The transition list was divided in half to achieve a minimum dwell time of 50 ms and each timepoint sample analysed with both transition lists. The digestion progress was monitored by comparing peak areas between the isotopic variants.

The "double-digestion" study used a nanoAcquity UPLC™ coupled to a Synapt™ G2 mass spectrometer (Waters) and acquired data using a MS^E program with 1 s scan times and a collision energy ramp of 15–40 eV for elevated energy scans. The mass spectrometer was calibrated before use against the fragment ions of glufibrinopeptide and throughout the analytical run at 1 min intervals using the NanoLockSpray™ source with glufibrinopeptide. Peptide identification was performed by using ProteinLynx Global SERVER™ v2.4 (Waters) to search a custom database of approximately 50 QconCATS. The data was processed using a low energy threshold of 100 and an elevated energy threshold of 20. A fixed carbamidomethyl modification for cysteine and isotopically labelled lysine and arginine were specified. The search thresholds used were: minimum fragment ion matches per peptide 3; minimum fragment ion matches per protein 7; minimum peptides per protein 1 and a false positive value of 4. Quantification was performed by creating extracted ion chromatograms of the most dominant charge state in the low energy data channel.

3. Results and discussion

3.1. Incomplete proteolysis in a quantification reaction

An example of the difficulty posed by incomplete proteolysis in absolute quantification is provided as part of a programme to quantify the yeast proteome [13]. In this instance, a high

abundance molecular chaperone (HSP12) was quantified with two different peptides, but the quantification, in copies per cell, differed between the two peptides. One peptide yielded a quantification result of about 900,000 copies per cell, but the second gave a value about half as high. Because all peptides were selected for uniqueness in the proteome, it was more plausible to suggest that the second peptide was yielding a low signal because of partial post-translational modification of the analyte peptide sequence, or because of incomplete excision of the peptide from the parent protein. Closer examination of the data revealed that the peptide that revealed the lower protein abundance (LNDAVEYVSGR) was partially represented in a mis-cleaved sequence (**SK**LNDAVEY VSGR/), itself derived by tryptic cleavage of /DYMGAAK/SKLNDAVE YVSGR/ (Fig. 2).

The selection of quantotypic peptides must therefore address the local sequence context beyond the immediate cleavage sites that generate the peptides. It is possible that the peptide **SK**LNDAVEYVSGR and LNDAVEYVSGR represent parallel but 'dead-end' proteolytic processes, such that diversion of the reaction to **SK**LNDAVEYVSGR would impair quantification. There is little evidence for trypsin having the capacity to function as a dipeptidyl peptidase (removing dipeptides from the N-terminus, thus removing **SK** from **SK**LNDAVEYVSGR) reinforcing this conjecture. In this instance, prolongation or enhancement of the proteolytic reaction might never achieve complete digestion, and arguably, the mis-cleaved peptide generated by one branch of the cleavage pathway is effectively a limit peptide (Fig. 3). By contrast, trypsin seems able to act as a peptidyl dipeptidase (removing dipeptides from a free C-terminus). If trypsin was more able to function in removal of dipeptides from the C-terminus, we might expect the nature of mis-cleaved products to vary depending on whether the adjunct fragment was N-terminal or C-terminal to the sibling peptide.

To explore this further, we analysed all mis-cleaved peptides in a complete trypsin digestion of a yeast extract, analysed by ion-mobility enhanced data independent mass spectrometry (HDMS^E). In total we were able to acquire data for 1415 ± 88 mis-cleaved peptides over four biological replicates, each independently digested under identical conditions. Of these, approximately 51% were N-terminal mis-cleaved peptides, containing short extensions at the N-terminus (up to five amino acids) and 32% contained C-terminal extensions – the remainder comprised larger peptide lengths on either side of the scissile bond (Fig. 3). The frequency of this last class is very similar at both termini.

The only sources of single amino acid mis-cleaved peptides (Fig. 4, $n = 1$) should be dibasic sequences, terminal peptides or neo-terminal peptides caused by internal proteolytic events prior to tryptic digestion. Examination of the data revealed that the mis-cleaved peptides that contained a single amino acid either N or C-terminal to the remainder of the peptide are almost exclusively the outcome of dibasic cleavage sites (the remainder being true N and C-termini). A small percentage of these lack a C-terminal basic residue because they are derived from the C-terminus of the protein. Interestingly, there is an equal preponderance of BxxxxxB or xxxxxxBB (where B is a basic residue) implying very little bias in cleavage of the two tryptic sites in a dibasic sequence. By implication, it is difficult for trypsin to 'clean' these mis-cleaved peptides by acting as an aminopeptidase or a carboxypeptidase. These are likely to be 'dead-end' products.

For the mis-cleaved peptides (Fig. 4, $n = 2$) that contain an extension of two amino acids, there is a pronounced bias to N-terminal extensions. Given the lack of bias in trypsin selectivity at dibasic sequences, we suggest that this bias reflects post-processing of the peptides, and that trypsin is more able to act as peptidyl dipeptidase (acting at the C-terminus) than a dipeptidyl peptidase (acting at the N-terminus). This may then explain the persistence of the peptide **SK**LNDAVEYVSGR in the example above.

Arginine and lysine are represented in the *S. cerevisiae* proteome at 4.4% and 7.2% of the total amino acid content respectively and thus, the frequency of the interspersed dibasic sequence [B]x[B] (where x is none of K,R and P) is approximately 1 in 100. In a typical protein of 300 amino acids, we might expect to encounter three such interspersed dibasic sites, each of which has the potential to compromise complete cleavage of two peptides flanking the site. A search of a UniProt *S. cerevisiae* protein database (6765 sequences) revealed 372,000 basic residues, ~15,000 of which were adjacent to a proline residue and thus, generally considered

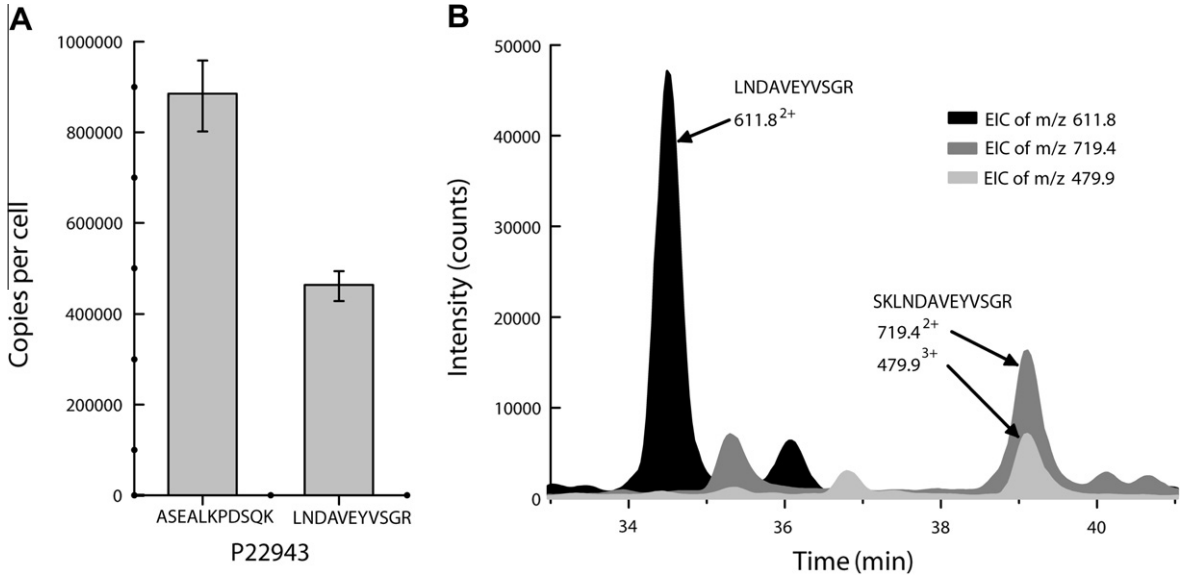

Fig. 2. Errors in quantification introduced by incomplete proteolytic digestion. The yeast protein HSP12 (P22943) was quantified by QconCAT methodology. Good practice recommends that protein quantification should be achieved by multiple peptides. In this case, the two peptides produced different abundance values for the protein ((A) error bars represent SEM of biological variation, $n = 5$). Further investigation of the tryptic digest reveals in addition to the limit peptide LNDAVEYVSGR ($[M+2H]^{2+}$: m/z 611.8), the presence of the mis-cleaved peptide SKLNDAVEYVSGR (present in both doubly) ($[M+2H]^{2+}$: m/z 719.4) and triply ($[M+3H]^{3+}$: m/z 479.9) charged forms. If one assumes a similar, or slightly reduced, response factor for SKLNDAVEYVSGR then the amount of limit peptide locked in this mis-cleaved product could account for the differences between the quantifications (B). EIC: extracted ion chromatogram.

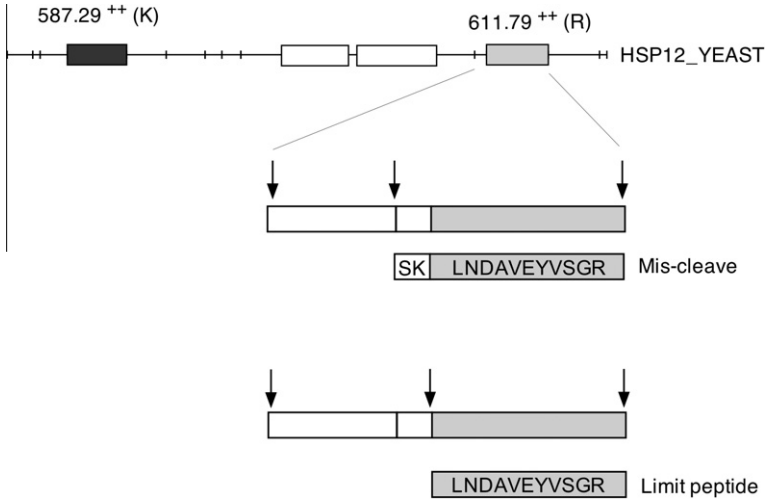

Fig. 3. Alternative proteolytic pathways can compromise quantification. For the quantification of yeast HSP12, two peptides were selected as candidates for quantification. One ($[M+2H]^{2+}$: m/z 587.3) cleaves from the protein cleanly and completely. The second ($[M+2H]^{2+}$: m/z 611.8) is in an interspersed dibasic context, leading to two distinct outcomes that are both 'dead-end' or limit products. In this instance, the failure to completely excise the true limited peptide has compromised the quantification (Fig. 2).

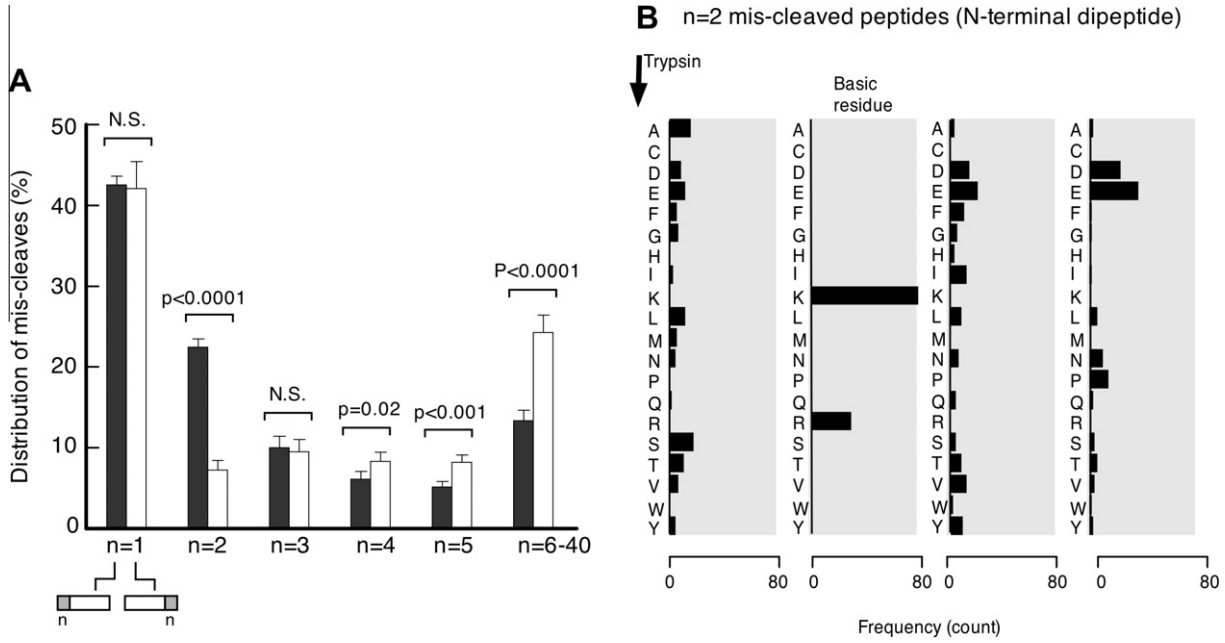

Fig. 4. Analysis of mis-cleaved peptides. Four biological replicates of a *Saccharomyces cerevisiae* proteome preparation were digested under identical conditions. The digest was analysed by HDMSE methodology on a Waters Synapt G2 instrument. From the peptides obtained, mis-cleaved peptides were extricated and recovered for the analysis. (Panel A) The distribution of the peptides, over the four biological replicates (errors bars are SD) were subdivided into those that had an extension at the N-terminus (closed bars) or at the C-terminus (open bars) and are distributed according to the length of the extension. Differences between the frequency of N and C-terminal extension sequences were assessed using an unpaired t-test. (Panel B) For the dipeptide N-terminal extension set, we analysed 105 such mis-cleaved peptides for the distribution of amino acids in the surrounding positions. The preponderance of acidic residues in the P1′ and P2′ positions reinforces the likelihood that it is the presence of such disfavoured residues that directs the preferred hydrolysis towards the mis-cleaved, dipeptide extended product, in accordance with the predictions of [4].

inaccessible. Searching for BB, BxB, BxxB and BxxxB (B = R or K) revealed a total of ~165,000 such motifs approximately equally distributed through the different sequences. For the sequence BxB, ~40,000 such sequences comprised ~1600 BPB and 7100 BBB, the remainder having the potential to generate an N-terminal or a C-terminal dipeptide extension. On average, a total of five to six peptides from each protein are compromised – a substantial loss of potential quantotypic peptides in quantitative studies. We note in passing that many proteotypic peptides are the product of cleavage at dibasic sites and the distributions of the peptide into the different products have yet to be rigorously explored. Such a study would also need to be cognizant of the different propensity of the peptides to ionize and fragment.

3.2. Variation in kinetics of release of standard and analyte peptides

Exploratory studies to establish optimal conditions for digestion should assess the rate of release and extent of release of the specific limit peptides of interest. In proteolytic reactions, the rate of digestion is most simply defined as a pseudo-first order rate constant (the protease is not consumed by the reaction). This is most accurately obtained by monitoring of a time course of digestion, whereby regular samples are taken from a digest over time (t) and the disappearance of substrate or formation of product peptides determined. The pseudo-first order rate constant (k) is obtained by nonlinear curve fitting of the ([*peptide*], t) data. In rapid proteolytic reactions, the product achieves a plateau value

very quickly, and provided sufficient data points are obtained in the pre-plateau phase of the reaction, the rate constant can be recovered. Moreover, a rapid reaction minimises the influence of nonspecific degradation events [14,15].

There are three types of behaviour that define the time course of proteolysis in a QconCAT experiment. First, the rate of release of the same peptide from standard and analyte can be similar, in which instance, reliable quantification will be obtained at around seven half times ($7 \times \ln(2)/k \approx 4.8/k$). Thus a reaction with a first order rate constant of 1.0 h^{-1} (a half time of about 40 min) should be essentially complete at 5 h. However, if the QconCAT or analyte are digested at dissimilar rates, then the reaction time required is defined by the slower of the two reactions. As an example, the proteolytic release of peptides from a QconCAT and analyte are presented in Fig. 5.

The first peptide, VTPSFVAFTPEER, exhibits ideal behaviour, inasmuch as both standard and analyte are fully released well within the reaction time. The rates of appearance of the peptides are about threefold different (standard faster than analyte) but both reactions are rapid. The second peptide, LVTGVNPASAHSTAVR is released 20-fold more rapidly from analyte than standard but both attain complete (plateau) digestion. The third peptide AIDLVDEACAVLR, exhibits the opposite behaviour, inasmuch as the standard peptide is released about 50-fold more rapidly than the same peptide from the analyte. A final peptide, DILGDVEQK represents the worst case scenario inasmuch as neither standard nor analyte are hydrolysed rapidly and there is no reason to assume that the proteolytic reaction is complete or that the ensuing quantification is reliable until the reaction has progressed to the plateau for both reactions.

It is most likely that the differential rates of hydrolysis reflect the nature of the amino acids that flank the scissile bond. In these examples, low rates of digestion are associated with acidic residues in P1′ or P2 positions that are known to impair efficient proteolysis [4]. Similar behaviours have been observed by Proc et al. [16]. Other than avoiding acidic residues in these positions, there are few rules that can be used to ensure full digestion of an analyte (and standard, if a QconCAT is used) and only the most obvious can be implemented [13]. We note in passing that the interdigitation of each quantotypic peptide with short intervening sequences [17] does not resolve this problem of incomplete proteolysis if the acidic residues are inherent to the analyte peptide, or are contained within the peptide in positions that disfavour complete hydrolysis. In our view, the only effective solution is to establish digestion

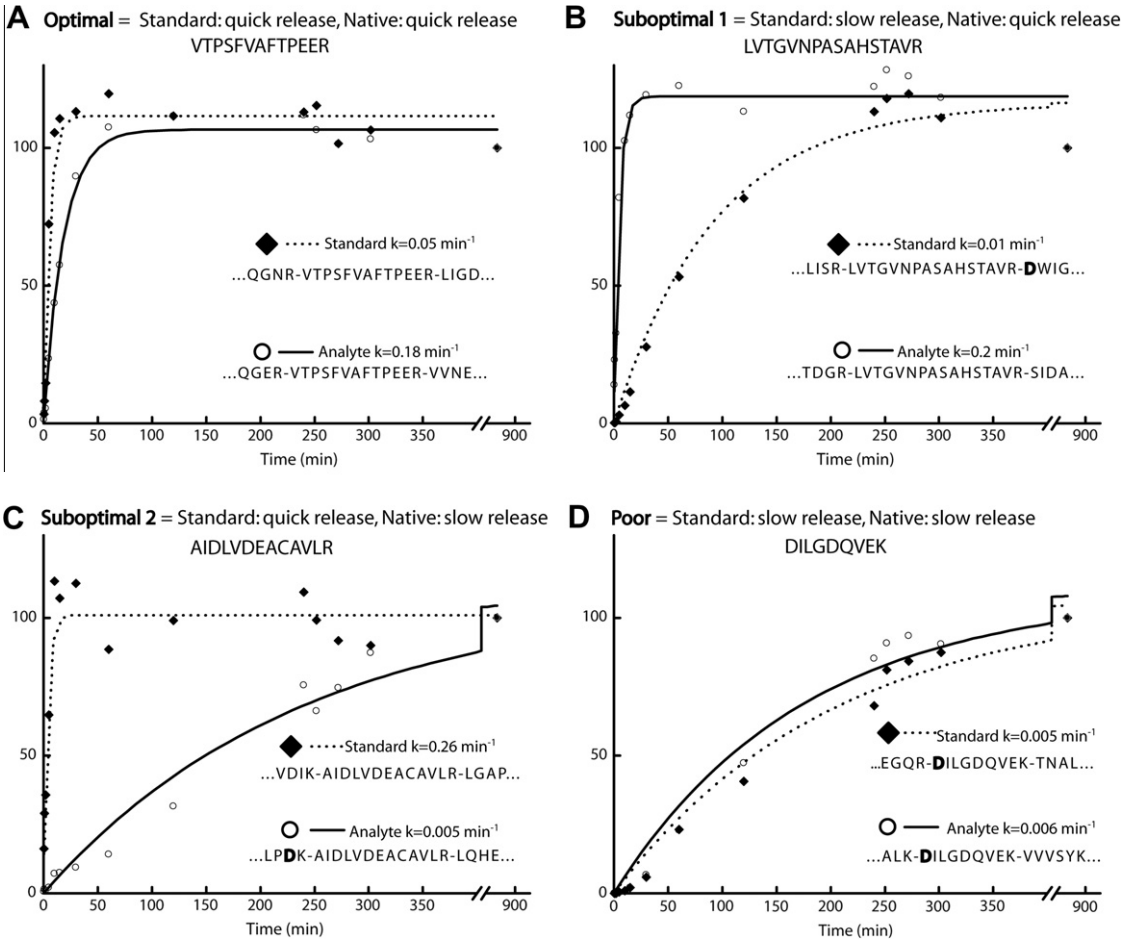

Fig. 5. Time course analysis showing the release of four peptides during tryptic digestion. A QconCAT standard was codigested with yeast and the mixture sampled over time and analysed by LC-SRM mass spectrometry. The QconCAT was isotopically labelled with [^{13}C6] lysine and [$^{13}C_6$] arginine allowing the release of the peptide from the QconCAT (dotted line) and native protein (solid line) to be followed. The first order rate constants (k) for each peptide are shown. The sequences flanking the digestion site are shown with residues that are known to be detrimental to digestion [4] highlighted in bold type. Optimum quantification will be produced in cases where release is rapid from both QconCAT and native protein (A). In cases where the release rate is different between QconCAT and native protein (B and C), quantification errors can occur because the two peptides are subjected to different environments. For cases where digestion occurs slowly (D) it is important to maintain constant digestion times in order to obtain reproducible quantification results. In these examples the reduction in digestion rate occurs because of the presence of an aspartate residue in close proximity to the digestion site. All peptides were quantified for the full reaction time of 900 min, and all data are expressed as a percentage of the peptide signal intensity measured at 900 min, set therefore to the common datum (900 min, 100%), where the symbols for each peptide overlap fully.

conditions that permit full proteolysis, irrespective of the flanking amino acid identity and to formally test for this completeness of digestion.

3.3. Optimal digestion for quantitative proteomics

With quantitative proteomics digestion, optimisation of proteolysis must be assessed by completion of digestion. Digestion protocols must be judged not just by the success of identification but also by reducing the number of mis-cleaved peptides. This focus on the "end game" of proteolysis where digestion has progressed virtually to completion emphasises peptides that contain a single mis-cleave site. At this stage, higher order protein structure will no longer be a determinant of proteolysis, and the digestion rate will be dictated predominantly by the primary structure of the peptide. Analysis of mis-cleaved peptides can be challenging; they will tend to be larger peptides and so are difficult to identify due to poor fragmentation with CID and poorer elution profiles. Assuming that the digestion protocol is effective they will also be in low abundance. In our analyses peptides with mis-cleaved sites on average yield lower scores from search engines. Although it is possible to monitor the enzymatic digestion in real-time [18], this approach focuses on the initial phase of digestion, where the rate of peptide production is high, as opposed to the detection of mis-cleave peptides in the end stages of digestion.

3.4. Relationship between observable signal intensities of mis-cleaved and limit peptides

One approach to study the properties of mis-cleaved tryptic peptides is to use the endoprotease Lys-C to deliberately create tryptic "mis-cleaved" peptides in which all Arg_X bonds remain unhydrolysed (Fig. 6). These "mis-cleaved" peptides can then be converted to the tryptic limit peptides by digestion with trypsin. By reducing the reaction mixture to a singly mis-cleaved precursor and two products, we are able to establish the response factor relationship between a mis-cleave precursor peptide and its product peptides (Fig. 7). It is clear that in many cases (Fig. 7A–D) the response factor of the mis-cleaved peptide is substantially lower than its descendant limit peptides. Thus, a low mis-cleaved peptide signal might not be considered to be indicative of significantly incomplete digestion but could in fact conceal the majority of the signal that should have been apparent in the limit peptides. Thus, identification and amelioration of mis-cleaved species is critical in quantitative studies. To overcome the poor performance associated with mis-cleaved peptides it may be possible to apply targeted methods to identify them with high sensitivity; Norrgran et al. included SRM transitions targeting mis-cleaved forms of their quantification peptide in their experimental strategy [19]. This approach is ideal for low complexity analyses but the increased preparation and analytical time would make it impractical for large-scale proteomic applications.

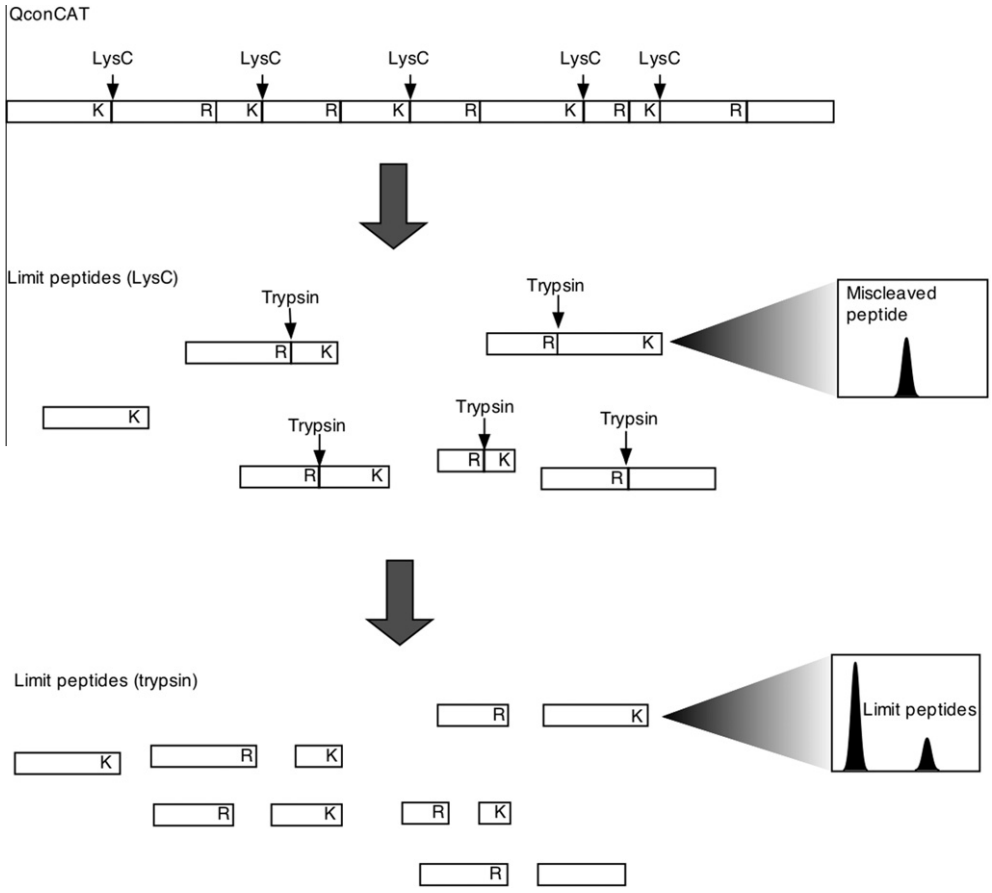

Fig. 6. Experimental protocol for comparing the response factors of a mis-cleaved peptide and its cognate limit peptides. A protein is digested with endoproteinase LysC to produce a mixture of peptides, some of which correspond to tryptic mis-cleaved peptides (XXXXRXXXK). These peptides are then tryptically digested to yield tryptic limit peptides. Comparison of the intensities of the LysC peptide and resultant tryptic peptides allow the response factors of the peptides to be compared. If the reaction is sampled over time it is possible to follow the rate of digestion at a single proteolysis site without the influence of higher order protein structure allowing the influence of primary sequence to be determined.

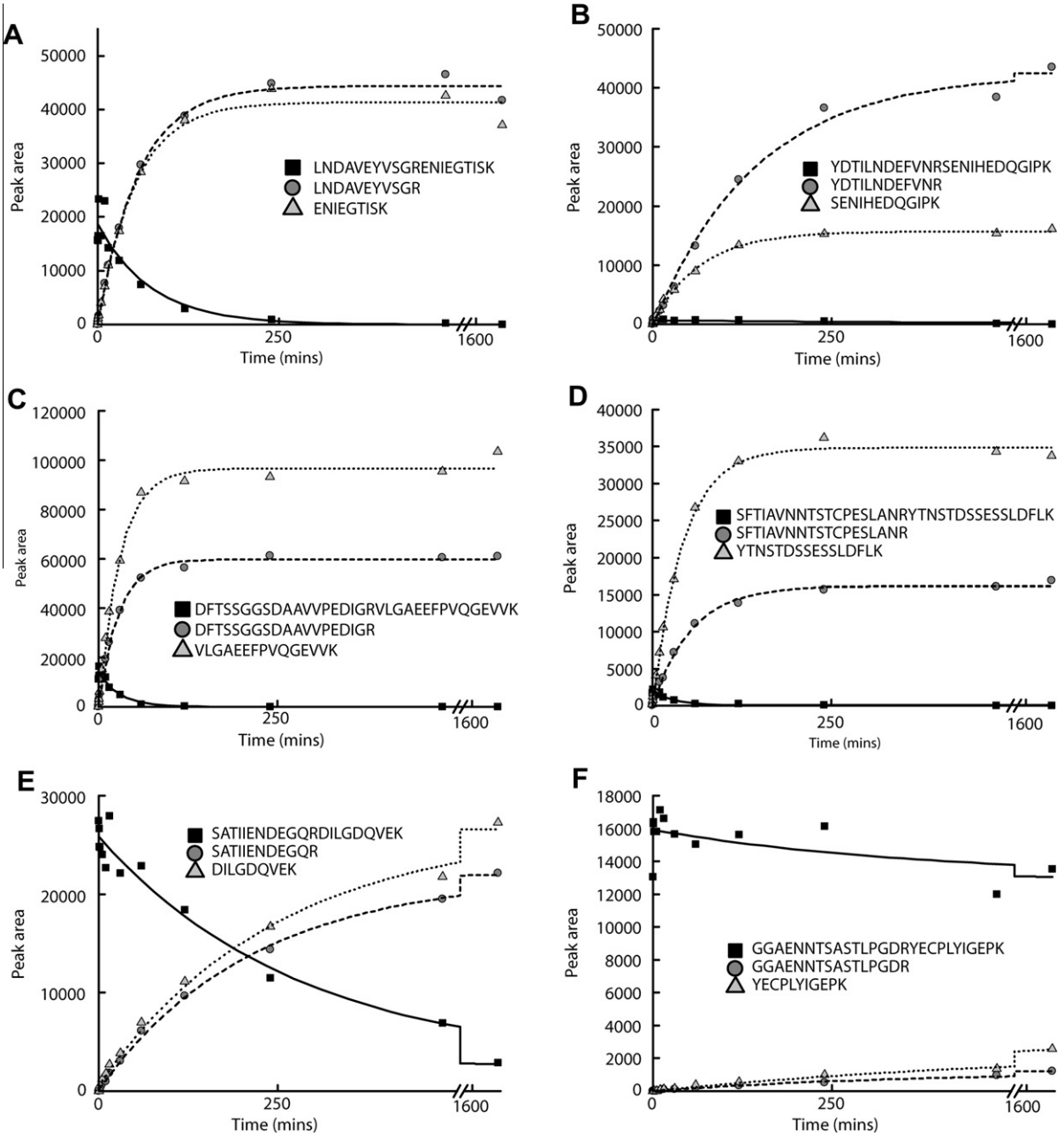

Fig. 7. The relative response factor of a mis-cleaved peptide in comparison to its digested cognate products. A mix of artificial QconCAT proteins was sequentially digested with LysC and trypsin. The tryptic digestion was sampled at regular intervals and analysed by LC–MS. The digestion of the LysC peptide (filled square) into the two tryptic peptides (open triangle and open circle) was quantified by determining the peak areas for each of the three peptides using extracted ion chromatograms. Fig. 7A–D demonstrate the hidden quantification danger of mis-cleavages in that the response factor of the mis-cleaved peptide is much lower than its product peptides. This means that a small mis-cleaved signal can actually indicate that the majority of a peptide may be locked in a mis-cleaved product. This experiment also allows the kinetics of single bonds to be determined in the absence of higher order protein structure allowing the influence of primary sequence to be investigated. Fig. 7E and F demonstrate the impact of acidic residues in vicinity of the cleavage site. The rate of digestion is reduced and in the case of GGAENNTSASTLPGDRYECPLYIGEPK, which contains the most hindering combination of aspartate in P1 and glutamate in P2′, virtually no digestion. The rate of DILGDQVEK release from SATIIENDEGQRDILGDQVEK is very similar to the rate of DILGDQVEK release from the intact QconCAT (Fig. 7D), 0.004 min^{-1} and 0.005 min^{-1} correspondingly, indicating the digestion of this site is the rate limiting step in the release of DILGDQVEK.

3.5. Strategies to ensure completeness of proteolysis

There have been many studies on the optimisation of trypsin proteolysis through modification of the parameters of digestion or the use of additives to enhance digestion. Earlier reports focused on qualitative improvements defined as the number of proteins identified or the sequence coverage achieved. As the focus of proteomics has shifted to quantification, recent studies have investigated the optimisation of digestion from the perspective of quantitative generation of peptides. Digestion enhancers destabilise higher order protein structure and improve protease access to unfolded polypeptide chains. A careful balance must be struck between destabilising the analyte protein structure and maintaining tryptic activity. Three main classes of destabilising additives can enhance digestion: organic solvents, chaotropes and surfactant. Solvents and chaotropes integrate well in a mass spectrometry workflow, it is only more recently that surfactants have been developed to be compatible with LC–MS. Organic solvents are presumed to function by destabilising the hydrophobic core of proteins, allowing access by the protease. Solvents tested include

(most commonly) acetonitrile, methanol and isopropanol up to concentrations of 80% (v/v), although lower concentrations are more commonly used [20]. A major appeal of organic solvent additives is the ease of removal; the solvent can be removed by evaporation or by dilution to a non-interfering concentration. Chaotropes destabilise higher order protein structure by disrupting the hydrogen bonding network in proteins. Urea is the most commonly used chaotrope because it has excellent compatibility with proteases (LysC retains a majority of its activity at 4 M urea and trypsin at 1 M urea [21,22]) and can be incorporated into a proteomics workflow by either dilution or online/offline reverse-phase clean up. For quantitative proteomics urea is often avoided because of it can be converted to isocyanate which in turn can carbamylate the side chain of lysine residues; a chemical modification that would alter retention time and split the peptide signal. The other major chaotrope, guanidine, is rarely used as a digestion additive because trypsin only retains activity at low concentrations (ca. 0.1 M). Many surfactants are incompatible with reverse-phase peptide separation and often incompatible with mass spectrometry (either MALDI and electrospray ionisation). Acid-labile surfactants (ALS) can be integrated into a proteomic workflow because it is possible to remove them by an acidification step prior to LC–MS, a process in which a surfactant is cleaved into a small hydrophilic section and a hydrophobic chain. The hydrophobic chain is insoluble and precipitates, such that it can be removed by centrifugation. ALS have now been commercialized and are available from several manufacturers. There is a growing appreciation that when compared to other forms of denaturant acid-labile surfactants are the most successful both qualitatively and quantitatively [16,23–26]. In a comprehensive study of digestion enhancement, Proc et al. investigated the use of digestion additives quantitatively using isotopically-labelled peptide standards to follow the digestion of 46 proteins by measuring the amount of released peptide over time. They reported that although no single additive resulted in complete digestion of all substrates, another acid-labile surfactant, sodium deoxycholate, produced the highest average digestion efficiency (~80%). Other experimental parameters (for example, temperature or solvent composition) produce little extra benefit over the addition of an ALS. The only optimisation required is the determination of an ALS concentration that does not hinder protease activity; for example, RapiGest™ can be used at 0.1% (w/v) without affecting tryptic activity [27]. A recent publication has confirmed that RapiGest™ also enhances access and digestion of membrane proteins [28].

Digestion reactors allow the solutions around the proteins to be exchanged before digestion without loss of the protein. This can either be achieved by retaining the proteins on beads or by a size exclusion membrane. The latter has been employed in a method developed some years ago [29] but more recently revisited as FASP [30] and improves digestion by allowing the use of a strong MS incompatible detergent (SDS) in a proteomics workflow. The size exclusion spin filter allows the SDS to be removed and exchanged for a strong urea solution. Bead based methods can enhance digestion efficiency both in terms of an increased number of identified proteins and a reduction in the digestion time [31]. This increased efficiency has been attributed to a localised increase in enzyme: substrate ratio at the bead surface.

3.6. The importance of complete proteolysis in label-free proteome quantification

The increasing interest in label-free quantification is understandable, as it requires no added steps to generate stable isotope labelled comparators, simply being based on the MS signals of different peptides that are derived from a protein. The two approaches used most commonly are spectral counting (based on the number of instances of matching of peptides in the database search algorithm) [7] or intensity-based methods, in which the intensity of multiple ions derived from a single protein are summed [6]. In either methodology, it is not clear that the influence of mis-cleaved products have been formally addressed. Mis-cleaved peptides can increase the number of apparent matches in spectral counting, and can be included in intensity calculations. Alternatively, if the abundance calculations are restricted such as to exclude mis-cleaved peptides, the loss of signal from the limit peptide signal that is caused by incomplete digestion could compromise the quality of the label-free quantification. We would advocate that all identification proteomics, as well as quantification analyses should include a specific index of digestion efficiency, similar to that suggested by Stead et al. [32]. However, the complex relationship between ion currents of mis-cleaved precursors and their cognate products precludes a simple statistic, and there is scope for new approaches to derive a figure of merit for digestion efficiency. This statistic can only be derived after database searching is complete, since it is the identification of the protein that establishes the relationship between limit peptides and mis-cleaved peptides.

4. Conclusion

Given the role of proteolysis (particularly trypsin) in bottom-up proteomics, it is surprising that there should be uncertainty about the nature of a tryptic digestion of a proteome. There is likely to be considerable variance between a predicted tryptic digestion (in which every bond is cleaved *in silico*) and an actual digestion mixture, even when the digestion reaction has reached completion. The complexity of the proteolytic reaction pathway could have the consequence that it is not fully reproducible, whether within or between laboratories, which might result in different label-free quantification data. Similarly, the selection of AQUA or QconCAT peptides might be directed to avoid interspersed dibasic or indeed, simple dibasic contexts, which may reduce the number of peptides that could be selected.

Acknowledgments

This work is supported by a grant from the Biotechnology and Biological Sciences Research Council to RJB (BB/G009112/1). We are grateful to Dr. Karin Lanthaler for the supply of the yeast broken cell preparation, and to Victoria Harman for expression and purification of the QconCATs. The QconCATs were designed by Dr. Craig Lawless as part of the COPY project and we are grateful for his input.

References

[1] K. Linderstrøm-Lang, Proteins and Enzymes, Stanford University Press, Stanford, 1952.
[2] J.J. Perona, C.S. Craik, J. Biol. Chem. 272 (1997) 29987–29990.
[3] I. Schechter, A. Berger, Biochem. Biophys. Res. Co. 27 (1967) 157–162.
[4] J.A. Siepen, E.J. Keevil, D. Knight, S.J. Hubbard, J. Proteome Res. 6 (2007) 399–408.
[5] S.A. Gerber, J. Rush, O. Stemman, M.W. Kirschner, S.P. Gygi, Proc. Natl. Acad. Sci. USA 100 (2003) 6940–6945.
[6] J.C. Silva, R. Denny, C.A. Dorschel, M. Gorenstein, I.J. Kass, G.Z. Li, T. McKenna, M.J. Nold, K. Richardson, P. Young, S. Geromanos, Anal. Chem. 77 (2005) 2187–2200.
[7] Y. Ishihama, Y. Oda, T. Tabata, T. Sato, T. Nagasu, J. Rappsilber, M. Mann, Mol. Cell. Proteomics 4 (2005) 1265–1272.
[8] R.J. Beynon, M.K. Doherty, J.M. Pratt, S.J. Gaskell, Nat. Methods 2 (2005) 587–589.
[9] H. Johnson, S.C. Wong, D.M. Simpson, R.J. Beynon, S.J. Gaskell, J. Am. Soc. Mass Spectrom. 19 (2008) 973–977.
[10] J.M. Pratt, D.M. Simpson, M.K. Doherty, J. Rivers, S.J. Gaskell, R.J. Beynon, Nat. Protoc. 1 (2006) 1029–1043.
[11] J. Rivers, D.M. Simpson, D.H. Robertson, S.J. Gaskell, R.J. Beynon, Mol. Cell. Proteomics 6 (2007) 1416–1427.

[12] V. Brun, A. Dupuis, A. Adrait, M. Marcellin, D. Thomas, M. Court, F. Vandenesch, J. Garin, Mol. Cell. Proteomics 6 (2007) 2139–2149.
[13] P.J. Brownridge, S.W. Holman, S.J. Gaskell, C.M. Grant, V.M. Harman, S.J. Hubbard, K. Lanthaler, C. Lawless, R. O'cualain, P. Sims, R. Watkins, R.J. Beynon, Proteomics, 2011, in press, doi:10.1002/pmic.201100039.
[14] W.I. Burkitt, C. Pritchard, C. Arsene, A. Henrion, D. Bunk, G. O'Connor, Anal. Biochem. 376 (2008) 242–251.
[15] C.G. Arsene, R. Ohlendorf, W. Burkitt, C. Pritchard, A. Henrion, G. O'Connor, D.M. Bunk, B. Guttler, Anal. Chem. 80 (2008) 4154–4160.
[16] J.L. Proc, M.A. Kuzyk, D.B. Hardie, J. Yang, D.S. Smith, A.M. Jackson, C.E. Parker, C.H. Borchers, J. Proteome Res. 9 (2010) 5422–5437.
[17] K. Kito, K. Ota, T. Fujita, T. Ito, J. Proteome Res. 6 (2007) 792–800.
[18] P. Karuso, A.S. Crawford, D.A. Veal, G.B. Scott, H.Y. Choi, J. Proteome Res. 7 (2008) 361–366.
[19] J. Norrgran, T.L. Williams, A.R. Woolfitt, M.I. Solano, J.L. Pirkle, J.R. Barr, Anal. Biochem. 393 (2009) 48–55.
[20] W.K. Russell, Z.Y. Park, D.H. Russell, Anal. Chem. 73 (2001) 2682–2685.
[21] Roche Applied Science, Trypsin Sequencing Grade from Bovine Pancreas: Instructions for Use, 2009.
[22] Roche Applied Science, Endoproteinase Lys-C Sequencing Grade from *Lysobacter enzymogenes*: Instructions for Use, 2007.
[23] E.I. Chen, D. Cociorva, J.L. Norris, J.R. Yates 3rd, J. Proteome Res. 6 (2007) 2529–2538.
[24] W.J. Hervey, M.B. Strader, G.B. Hurst, J. Proteome Res. 6 (2007) 3054–3061.
[25] A.A. Klammer, M.J. MacCoss, J. Proteome Res. 5 (2006) 695–700.
[26] Y. Lin, J. Zhou, D. Bi, P. Chen, X. Wang, S. Liang, Anal. Biochem. 377 (2008) 259–266.
[27] Y.Q. Yu, M. Gilar, P.J. Lee, E.S. Bouvier, J.C. Gebler, Anal. Chem. 75 (2003) 6023–6028.
[28] F. Mbeunkui, M.B. Goshe, Proteomics 11 (2011) 898–911.
[29] L.L. Manza, S.L. Stamer, A.J.L. Ham, S.G. Codreanu, D.C. Liebler, Proteomics 5 (2005) 1742–1745.
[30] J.R. Wisniewski, A. Zougman, N. Nagaraj, M. Mann, Nat. Methods 6 (2009) 359–360.
[31] M. Ethier, W.M. Hou, H.S. Duewel, D. Figeys, J. Proteome Res. 5 (2006) 2754–2759.
[32] D.A. Stead, A. Preece, A.J.P. Brown, Mol. Cell. Proteomics 5 (2006) 1205–1211.